Learning the Power of "No": Foundation Models with Negations

Jaisidh Singh[1]*, Ishaan Shrivastava[2]*, Mayank Vatsa[3], Richa Singh[3], Aparna Bharati[4]
[1]University of Tübingen, Germany [2]Metafusion, India [3]IIT Jodhpur, India [4]Lehigh University, USA
jaisidh.singh@student.uni-tuebingen.de ishaan@metafusion.ai
{mvatsa, richa}@iitj.ac.in apb220@lehigh.edu

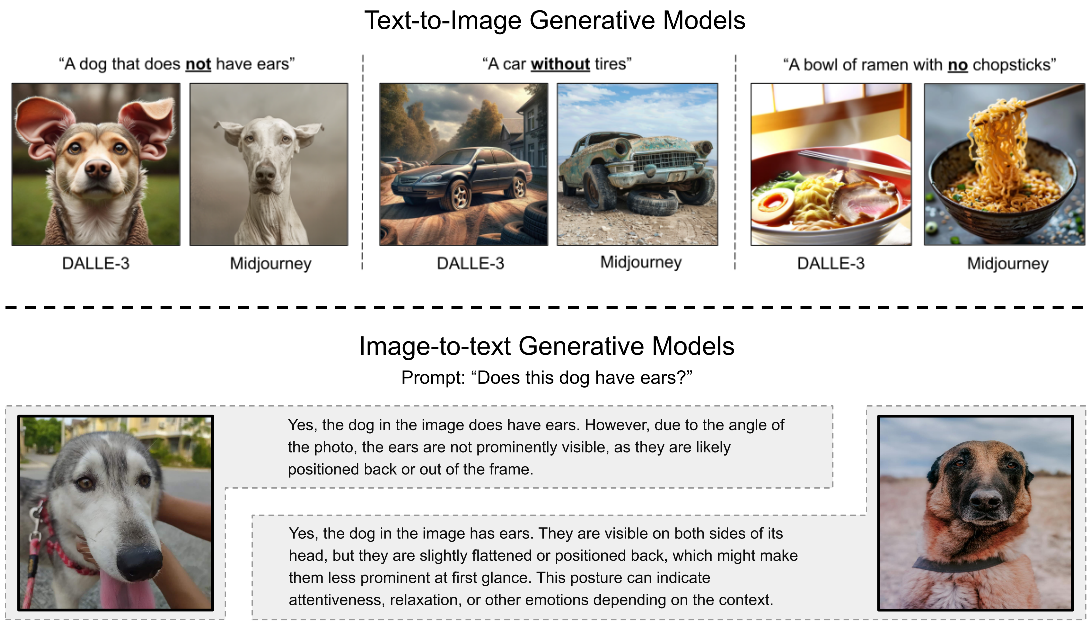

Figure 1. Vision-language models (VLMs) are utilized in multi-modal generative applications such as text-to-image generation (e.g., DALLE-3, Midjourney) and image-to-text generation (e.g., ChatGPT-4o). Above examples highlight the implications of lack of negation understanding in foundation models and motivate our work.

Abstract

Negation is a fundamental aspect of natural language reasoning, yet foundational vision-language models (VLMs) like CLIP face significant challenges in accurately interpreting it. These models often process text prompts holistically, making it difficult to isolate and understand the role of negated terms. To overcome this limitation, we present CC-Neg: a novel dataset consisting of 228,246 images, each paired with both true captions and their corresponding negated versions. CC-Neg provides a critical benchmark to assess and improve foundational VLMs' ability to process negations, focusing specifically on how the presence of terms like 'not' alters the semantic relationship between images and their textual descriptions. To illustrate the effectiveness of the CC-Neg dataset in enhancing negation understanding, we introduce the CoN-CLIP framework, which incorporates targeted modifications to CLIP's contrastive loss function. When trained with CC-Neg, CoN-CLIP achieves a 3.85% average improvement in top-1 accuracy for zero-shot image classification across eight datasets, and a 4.4% performance boost on challenging compositionality benchmarks such as SugarCREPE. These results highlight CoN-CLIP's enhanced understanding of the nuanced semantic relationships involving negation. Our code and the CC-Neg benchmark are available at: `https://github.com/jaisidhsingh/CoN-CLIP`.

*These authors contribute equally.

1. Introduction

Achieving generalized vision-language understanding is crucial for building high-performing multimodal foundation models [7, 17, 24–26, 30, 41, 48]. Contrastive learning is a powerful method for creating joint multimodal embedding spaces. It aligns representations of related images and texts while separating unrelated pairs based on semantic and visual similarities [24, 41]. Further, vision-language models (VLMs) [1, 24, 41, 48, 53, 56] are pretrained on large-scale image-text datasets [5, 47]. This enables them to excel in zero-shot tasks like image-text matching, image retrieval, and object classification. However, controlling the invariance learned by these models is difficult. Their generalization depends heavily on the quality and diversity of the training data, which affects adaptation to unseen contexts [44]. Additionally, the contrastive learning objective is optimized for retrieval tasks which can lead to "shortcut learning," where models behave like bag-of-words systems. As a result, they may have a limited understanding of relational semantics between concepts [55].

VLMs like CLIP [41] often ignore negation words such as *no, not,* and *without*. For example, an image of a dog matches with similar scores to both "*this is a photo of a dog*" and "*this is **not** a photo of a dog*" (Fig. 2). Further, Fig. 2 illustrates further that VLMs inadequately capture negation words, indicating under-representation in the training data and a misalignment of negations with their correct implications in the image space. Negations allow us to specify the absence of concepts [20] and hence form an important part of logic and natural language. However, negative sentences are harder to process than affirmative sentences [12, 38]. This is also highlighted through the under-representation of negatives in existing natural language inference benchmarks [15, 45] and that pretrained language models have difficulty performing well during neural translation tasks [14] and fill-in-the-blank tests [19]. Understanding negations, though harder for learning-based models [10, 49], is crucial for commonsense reasoning tasks [45, 46]. This ability is highly desirable in image-text retrieval and text-to-image generation systems [43].

To investigate this issue, we develop a comprehensive benchmark to evaluate VLMs' ability to understand explicit negations. We introduce the CC-Neg dataset, containing $228,246$ image-caption pairs accompanied by grammatically correct and fluent *negated captions*. The negated caption is a distractor text where a concept present in the image is negated explicitly using words such as *no, not,* and *without*. We use the CC-3M dataset to generate (*image, caption, negated caption*) triplets to test the negation understanding capabilities of VLMs. Through experiments on CC-Neg, we establish that VLMs generally do not understand prompts with negations and often match negated captions to the image over their true captions.

To mitigate this problem, we propose to augment the InfoNCE contrastive loss [37] with a contrastive objective, by leveraging fluent and high-quality negated captions in CC-Neg and distractor images. CLIP's text encoder [41] is fine-tuned using the proposed objective, and the resulting CoN-CLIP model shows improvements on the negation-understanding task across varyingly complex negated captions. Additionally, we find that our approach improves overall compositional understanding and outperforms CLIP by 4.4% average R@1 on SugarCREPE, an unbiased benchmark for tasks such as replacing, adding, and swapping objects, attributes and relations in prompts. This emphasizes CoN-CLIP's ability to understand the semantic decomposition of scenes into objects and their association with various attributes and relations, without explicitly being trained with compositional prompts beyond negations. Further, CoN-CLIP achieves improvements in top-1 zero-shot image-classification accuracy across 8 datasets, namely ImageNet-1k [8], CIFAR-10 [22], CIFAR-100 [22], Caltech-101 [11], Food-101 [3], Flowers-102 [36], Oxford Pets [39], and Stanford Cars [21], with the highest improvement being 10.95% on CIFAR-100. The contributions of this paper are as follows:

1. We demonstrate that VLMs struggle with negations, often misaligning them with images. For robust investigation of this phenomenon, we introduce CC-Neg, a large-scale dataset of $228,246$ image-caption pairs with high-quality negated captions as distractors.
2. Leveraging CC-Neg's captions and distractor images, we present a fine-tuning framework, CoN-CLIP, that improves upon InfoNCE contrastive loss [37] and enhances negation comprehension in pretrained models.
3. CoN-CLIP demonstrates enhanced performance on zero-shot image classification task and general purpose compositionality benchmarks, indicating a deeper understanding of visual concepts and improved compositional reasoning capabilities.

2. Related Work

Contrastive Image-Text Pretraining: CLIP, one of the most popular VLMs, is contrastively pretrained on approximately 400M image-text pairs, and has emerged to be applicable for several tasks such as open-set attribute recognition [6] and object detection [33]. New additions to CLIP's recipe such as image captioning with contrastive pretraining and self-supervision have produced models like BLIP [24], BLIP2 [23], SLIP [34]. As a foundation model, CLIP has been applied in image synthesis [42, 43], video-summarization [35], and has been extended to modalities such as video [4] and audio [13].

Compositional Understanding: Towards compositional image-text matching, [18] presents a model to decompose

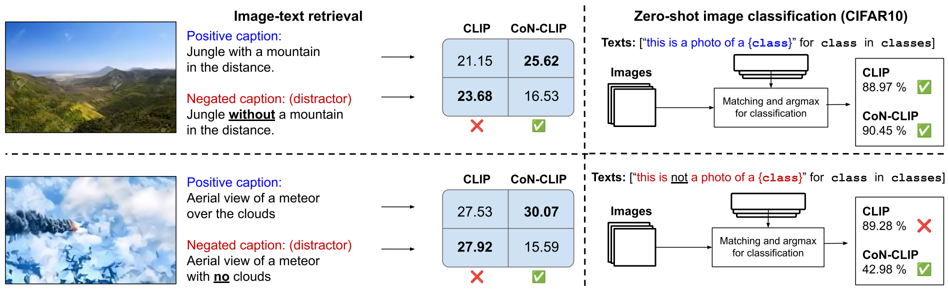

Figure 2. VLMs such as CLIP often match images to negation-based distractors with higher similarities than their true captions (left). Further, CLIP accurately retrieves images of a class even when prompted with "this is not a photo of a {`class`}" (right).

images and texts into respective sub-images and words denoting subjects, objects, and predicates. Along similar lines, [55] presents ARO, a benchmark to study the sensitivity of VLMs to object order, relations, and attributes. The study shows that VLMs struggle with compositionality, and presents NegCLIP to improve on the investigated shortcomings. Next, CREPE [31] presents a benchmark to evaluate compositionality in VLMs through systematicity and productivity. The systematicity component evaluates a VLM on seen and unseen contexts, while productivity entails image-text matching with various types of hard negative captions which act as distractors. SugarCREPE [16] refines biases in CREPE and ARO to present a high-quality unbaised dataset where Neg-CLIP shows significantly reduced improvements as compared to biased compositionality benchmarks like ARO and CREPE, implying an overfit on negative artifacts seen in training.

Using Hard Negatives and Negations: Hard negatives, or distractors which often lead to incorrect matching, are prominently used to evaluate image-text matching. CREPE and NegCLIP utilize such hard negatives to test sensitivity towards object order, swapping, relations, etc. Hard negatives are different from negations, which represent the absence of a concept. For instance, a simple negation is given by "this is *not* a cat", which implies an object that does belong to the cat class. CLIPN [52] devises a method to learn prompts for CLIP which correspond to "this is *not* X".

3. CC-Neg: Benchmarking Negation Understanding

Current datasets for image-text matching [5, 9, 27, 50, 51] largely focus on matching images to their true captions in the presence of distractors (either distractor images or distractor texts). However, negations are rare in such datasets. The Negate fold of CREPE-Productivity [31] is an example dataset, with 17K true image-caption pairs with 183K distractor texts. The distractors include negation words, but suffer in terms of linguistic fluency (Table. 1). This prevents the evaluation of negation understanding in VLMs in realistic settings. Hence, we introduce CC-Neg, a dataset aimed at comprehensively evaluating how well VLMs understand negations in realistic prompts.

Table 1. Comparison of CC-Neg with the Negate fold of CREPE-Productivity across true (*P*) and negated (N) caption pairs. CC-Neg contributes a larger scale and greater diversity in the type of negation words used. Further, it exhibits greater fluency and plausibility in its text data as indicated by higher mean Vera scores [16, 28] for the negated captions (0.347 for CC-Neg versus 0.232 for CREPE-Negate).

Dataset	**Captions *P* v/s N**
CREPE Negate	*P: Tree on a side of a street. street has on side a tree.* N: Tree on a side of a street. Street not has on side a tree.
	P: Car has tires. There is a windows. N: There is no car has tire. There is a windows.
CC-Neg	*P: Festive banner with flags and an inscription.* N: Festive banner with an inscription, but not with flags.
	P: A woman walks her dog on the beach. N: A woman walks on the beach without her dog.
	P: Dining table with kitchen in the background. N: Dining table with no kitchen.

3.1. CC-Neg Dataset

CC-Neg utilizes the Image-Labels subset, $300,000$ image-caption pairs, from the Conceptual Captions (CC-3M) dataset and a large language model (LLM) to obtain corresponding negated captions (overview in Fig. 3). Given an image-caption pair (I, c), we use PaLM-2 [2] to generate a negated caption c'. For example, a true caption such as "A city street with colorful billboards" is used to write a negated caption "A city street *without* billboards".

More specifically, the negated caption c' is obtained by prompting PaLM-2 to decompose c into one *subject*, and $\mathcal{K}$ *predicate-object pairs* using in-context learning (ICL) [32, 54]. Along with instructions to decompose the sentence into the above components, we add a handcrafted

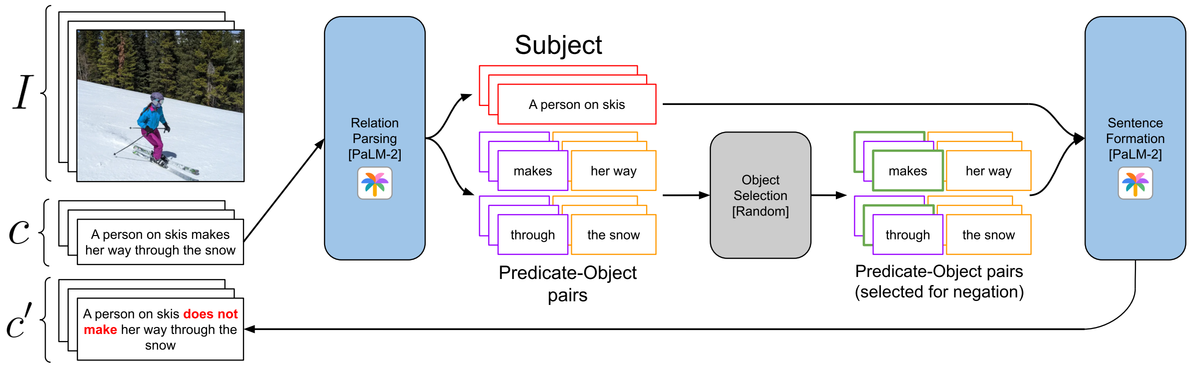

Figure 3. Overview of the generation of negated captions. Given the true caption of an image, an LLM (i) decomposes it into a *subject* and *predicate-object pairs*, and then (ii) selects a random predicate-object pair to negate to finally write the negated prompt.

example within the prompt as a demonstration of the task. This allows PaLM-2 to effectively follow the schema required for this task. More details regarding the prompting method can be found in Sec. A of the supplementary material. Next, for each caption, an object from the $\mathcal{K}$ pairs is randomly selected and its association to the subject as well as the scene is nullified using a negation word such as $\{not, no, without\}$ (Fig. 3). We use ICL in this step as well, where an example input and output is provided in the prompt as format. This results in the negated caption c'. In some cases, PaLM-2 does not faithfully decompose all captions. Additionally, PaLM-2 can negate objects by omission for certain samples. Such responses are considered erroneous and are excluded. Finally, CC-Neg contains $228,246$ (I, c, c') triplets. We use CC-Neg as the test data for 4 VLMs and evaluate their performance in associating true images with true captions in the following subsection.

3.2. Evaluating VLMs on CC-Neg

We benchmark four state-of-the-art VLMs: CLIP [41], BLIP [24], FLAVA [48], and Neg-CLIP [55] on CC-Neg, to test how well VLMs identify true image-caption pairing in the presence of negated captions as distractors.

Experimental setup: For each triplet (I, c, c') in CC-Neg, a VLM computes a similarity-based match score $\phi(\cdot, \cdot)$ between each image-text pair. If $\phi(I, c) > \phi(I, c')$, the VLM is deemed to match the image I to its true caption c over the distractor and indicates a correct prediction. Alternately, $\phi(I, c) \leq \phi(I, c')$ signifies an incorrect prediction. Using this rule, we compute the accuracy of identifying true pairings for each VLM. For fair comparison of CLIP with Neg-CLIP, we use the ViT-B/32 architecture for both models.

Results: The performance evaluation of state-of-the-art VLMs on CC-Neg results in the following observations.

1. **VLMs fail to recognize negations:** We find that all VLMs exhibit poor understanding of negations in text. The accuracy values in Table. 2 signify that VLMs often confuse negated captions as true ones. Specifically, the presence of the negated concept within I is erroneously associated with c', showing that VLMs largely ignore the effect of negation words like "not","without", etc. Notably, Neg-CLIP, which otherwise outperforms CLIP on the Negate fold of CREPE [55], does not show similar trends on CC-Neg. This can be attributed to our data generation procedure, where leveraging an LLM results in greater linguistic fluency in the negated captions. Consequently, our data domain differs from CREPE-Negate in the quality of distractor texts, which has more crude and non-fluent samples (shown in Table 1). This supports [16] in that Neg-CLIP exhibits biases towards non-fluent data. Overall, CLIP has the highest accuracy on CC-Neg, with Neg-CLIP, BLIP and FLAVA close but only slightly above random chance (50%).

Table 2. For each VLM, we report the model and pretraining configurations alongside its accuracy on the entire CC-Neg dataset.

Model	Architecture & Pretraining	CC-Neg Acc
CLIP	ViT-B/32 (OpenAI)	66.4
Neg-CLIP	ViT-B/32 (OpenAI+ARO fine-tuned)	62.7
FLAVA	Full (Meta)	60.8
BLIP	Base (Salesforce+COCO finetuned)	63.5

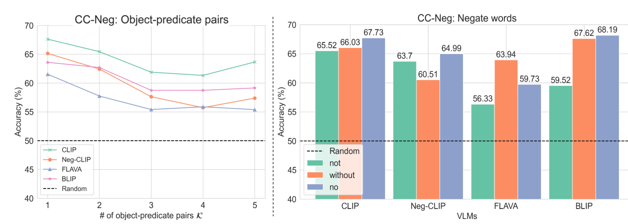

Figure 4. We report the accuracy of matching the image to its true caption for all VLMs, varying the number of predicate-objects, $\mathcal{K}$ from 1 to 5 (left). Additionally, we show the performance of all VLMs on each type of negation word used in CC-Neg (right).

2. **Performance degradation at higher complexities:** Next, we study the responses of the VLMs across all caption complexities (number of predicate-object pairs $\mathcal{K}$) in CC-Neg. Fig. 4 (left) depicts the accuracy of identifying true pairings for each value of $\mathcal{K}$. We find that models perform worse as the captions become more complex, arriving near random chance for all models except CLIP, supporting the claim that VLMs cannot compositionally understand negations. The presence of more objects and predicates likely obscures the effects of negation words and results in reduced performance.

3. **VLMs favor certain negation words:** Lastly, we evaluate the effect of each negation word on the accuracy of a VLM. Fig. 4 (right) reports the accuracy of matching true pairs when the negation word in c' is *not*, *without*, and *no*. CLIP, Neg-CLIP, and BLIP are most accurate on *no*, while FLAVA favors *without*, reflected in its lower *no* and *not* accuracies.

4. Compositional Understanding of Negations

To improve VLMs' understanding of negations, we use the CC-Neg dataset and present an improved contrastive CoN-CLIP framework. We incorporate negated captions and relevant distractor images for fine-tuning CLIP [41], in addition to the image and true caption pairs originally used.

CoN-CLIP aims to enable CLIP and similar VLMs to interpret negated captions and understand their impact on scene composition. This motivates two design choices:

1. **Negated captions per sample:** We utilize a subset of CC-Neg, our large-scale dataset containing negation-based distractor texts. Specifically, the negated caption c'_i is used alongside the true image-caption pair (I_i, c_i).

2. **Distractor images as reflections of negated captions:** Providing visual context has shown to help model negation and its implications [49]. To anchor the effect of negations to visual concepts, we add distractor images which serve as *crude* reflections of the negated caption c'. Repelling such a distractor image I' from the true caption c shall lead to improved compositional awareness.

   Given a true caption c and a negated caption c' from CC-Neg, we first segregate concepts present in the scene from those that are absent, depicted in c'. Specifically, we use the subject and the negated object obtained from the relation parsing output of PaLM-2 while generating c'. For a sample (I, c, c'), I' is selected by mining MSCOCO [27] for an image that (i) contains the subject of the true caption, and (ii) does not contain the negated object. For example, the distractor image corresponding to the caption "A building in the sunset" shall contain a building but not the sunset environment, in alignment with the negated caption. More details about this process are provided in Sec B of the supplementary material.

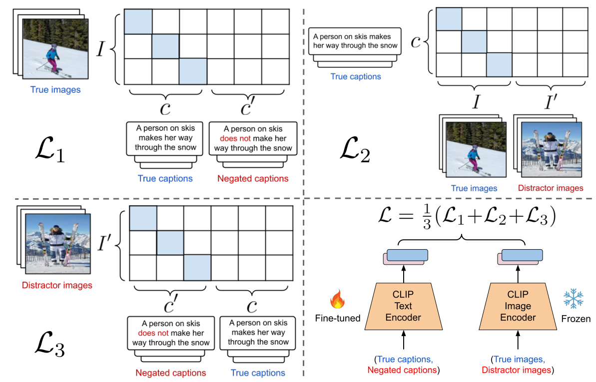

Figure 5. We incorporate negations and distractor images in a contrastive objective for fine-tuning the CLIP text encoder towards improved negation understanding.

Using negated captions and distractor images alongside the existing image-caption pairs, we compile a dataset $\mathcal{D} = \{I_i, c_i, c'_i, I'_i\}_{i=1}^{N}$. Here, N is set to $188,246$ to hold out the remaining $40,000$ samples in CC-Neg for evaluation. Next, we present the contrastive learning used in CoN-CLIP.

4.1. Fine-tuning CLIP with New Objectives

As shown in Fig. 5, the modification of the contrastive objective of CLIP is given as follows. Let $f_{img}(\cdot)$ denote the image encoder and $f_{txt}(\cdot)$ the text encoder of CLIP. Using these encoders, we embed a set of M images $\mathcal{I} = \{I_1, ...I_M\}$ and a set of M captions $\mathcal{C} = \{c_1, ..., c_M\}$ to $E_{\mathcal{I}}$ and $E_{\mathcal{C}}$ respectively. Similarly, a set of negated captions $\mathcal{C}' = \{c'_1, ..., c'_M\}$ and a set of distractor images $\mathcal{I}' = \{I'_1, ...I'_M\}$ are embedded with their respective encoders to obtain $E_{\mathcal{C}'}$ and $E_{\mathcal{I}'}$. Here, each set of CLIP embedding belongs to $\mathbb{R}^{M \times d}$. We then construct 3 similarity matrices to be used in the final objective. $E_{\mathcal{C}}$ and $E_{\mathcal{C}'}$ are concatenated and the cosine-similarity of the concatenated caption embeddings with $E_{\mathcal{I}}$ are computed to obtain $T_1 \in \mathbb{R}^{M \times 2M}$. $E_{\mathcal{I}}$ and $E'_{\mathcal{I}}$ are concatenated to compute their cosine-similarity with $E_{\mathcal{C}}$. The resultant similarity matrix is denoted by $T_2 \in \mathbb{R}^{M \times 2M}$. Lastly, $E_{\mathcal{C}'}$ and $E_{\mathcal{C}}$ are concatenated after which the cosine-similarity matrix between $E_{\mathcal{I}'}$ and the concatenated image embeddings is computed as $T_3 \in \mathbb{R}^{M \times 2M}$. The matrices T_1, T_2, T_3 are subsequently scaled by τ and column-wise softmaxed to give, $\tilde{T_1}, \tilde{T_2}, \tilde{T_3}$. This process, for any paired embedding sets $X \in \mathbb{R}^{N_1 \times D}$ and $Y \in \mathbb{R}^{N_2 \times D}$, is denoted by

$$(\tilde{T})_{ij} = \frac{e^{\tau X_i Y_j^T}}{\sum_{k=1}^{N_2} e^{\tau X_i Y_k^T}}.$$

$\tilde{T}_1, \tilde{T}_2, \tilde{T}_3$ are then used in the following formula to provide 3 loss terms $\mathcal{L}_1, \mathcal{L}_2,$ and $\mathcal{L}_3$ respectively.

$$\mathcal{L}_k = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{2M}\mathbb{1}_{\{i=j\}}\log((\tilde{T}_k)_{ij}) \tag{1}$$

Finally, we compute the total loss $\mathcal{L}_{conclip}$ as $\mathcal{L}_{conclip} = \frac{1}{3}(\mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3)$. Observing the lack of understanding of negations in text, it becomes necessary to train the embedding layer and the attention mechanisms of the text encoder. This is done to impart new knowledge of how negations affect the semantics of the given scene. Hence, we freeze the image encoder and fine-tune CLIP's text encoder on the final loss function $\mathcal{L}_{conclip}$, similar to [56]. The learning rate is initialized as, $1e{-}6$ which follows a cosine schedule of 50 warmup steps. The optimizer used is AdamW [29] with 0.2 weight decay and a batch of size 256. We use PyTorch [40] and run all experiments on one NVIDIA V100 GPU.

4.2. Experiments

This section outlines experiments assessing the proposed framework across various tasks and comparing it with existing methods. First, we evaluate CoN-CLIP's ability to understand negations (Sec. 4.2.1). Next, Sec. 4.2.2 explores the impact of CoN-CLIP's contrastive loss modifications on zero-shot image classification. Finally, Sec. 4.2.3 examines CoN-CLIP's compositionality capabilities. In these experiments, CoN-CLIP refers to CLIP fine-tuned on CC-Neg.

4.2.1 Understanding of Negations

CoN-CLIP is compared with other VLMs mentioned in Sec. 3.2 using the ViT-B/32 backbone across all CLIP-based models for fair comparison. To test our framework's understanding of negations, we use a held-out evaluation set from CC-Neg containing $40,000$ (I, c, c') triplets to measure the accuracy of matching image I to the true caption c in the presence of the negated caption c' as explained in Sec. 3.2.

Results: Matching accuracy of CoN-CLIP on CC-Neg evaluation set is reported in Table. 3 alongside CLIP, Neg-CLIP, FLAVA, and BLIP. CoN-CLIP outperforms other VLMs by a large margin ($> 30\%$) on the held-out samples for all caption complexities (number of predicate-object pairs $\mathcal{K}$). While CoN-CLIP's performance decreases as the value of $\mathcal{K}$ increases, the drop in performance is significantly less and does not fall below 99% even for $\mathcal{K} = 5$. Similarly, CoN-CLIP improves in performance across each type of negation word. Here, CoN-CLIP performs the worst for negated captions containing *no* as the negation word (96.5%), while still outperforming other VLMs (best being BLIP at 68.19%) on such samples. These results are provided in further detail in Sec. C of the supp. mat. These results show that CoN-CLIP exhibits a greater understanding of negations in text as compared to other VLMs. Further, it learns to reliably reject captions which negate visually-present concepts.

Table 3. Evaluating CoN-CLIP and other VLMs on CC-Neg. Underlined values denote highest performance across all models.

Model	Architecture & Pretraining	CC-Neg Acc ↑
CLIP	ViT-B/32 (OpenAI)	65.70
Neg-CLIP	ViT-B/32 (OpenAI+ARO fine-tuned)	62.63
FLAVA	Full (Meta)	58.93
BLIP	Base (Salesforce+COCO fine-tuned)	62.31
CoN-CLIP	ViT-B/32 (OpenAI+CC-Neg fine-tuned)	<u>99.70</u>

Additionally, we evaluate if CoN-CLIP can transfer its understanding to prompts which directly negate the subject of the text. For this, we use 8 popular benchmarks for image classification. Following the example in Fig. 2, image classification accuracy is computed using two types of class prompts: "this is a photo of a {class}" (standard), and "this is not a photo of a {class}" (negated). The latter must be matched to images which do not belong to the "class" category, indicated in low top-1 accuracy. To benchmark this behavior, we compute Δ, the difference between top-1 accuracies obtained by using standard class prompts and those obtained by using negated class prompts. This Δ value is computed for 8 image classification datasets, namely ImageNet-1k [8], CIFAR-10 [22], CIFAR-100 [22], Caltech-101 [11], Food-101 [3], Flowers-102 [36], Oxford Pets [39], and Stanford Cars [21], and averaged to obtain a single measure. It is desirable to show high accuracy while using standard class prompts, however, negated prompts for a given class must show low accuracy. CoN-CLIP is able to correctly reject images when observing negated class prompts indicated in the significantly higher mean Δ for CoN-CLIP (62.03%) versus that of CLIP (0.98%). This shows the ability of CoN-CLIP to generalize to subject negations and correctly identify concepts to reject beyond its training data.

4.2.2 Zero-shot Image Classification

The framework addresses limitations in understanding negations and their visual associations. We further explore how CoN-CLIP fine-tuning impacts CLIP's performance across diverse tasks, evaluating its efficacy in zero-shot image classification. We evaluate the effect of our fine-tuning process across all CLIP architectures ViT-B/16, ViT-B/32, and ViT-L/14 which are also used as baselines for compari-

Table 4. Evaluation of CoN-CLIP on zero-shot image classification shows improvements across all datasets. Here, highest accuracy values for a dataset are underlined, while highest accuracy values for a CLIP backbone are given in *italics*.

Model	ImageNet 1k	Caltech 101	Flowers 102	CIFAR 100	Food 101	Stanford Cars	Oxford Pets	CIFAR 10
CLIP								
ViT-B/16	68.35	82.56	64.14	53.54	86.89	61.68	81.82	88.23
ViT-B/32	63.36	81.50	60.50	55.18	81.15	58.33	80.08	88.97
ViT-L/14	75.51	81.80	72.42	65.95	92.10	74.64	88.06	91.40
CoN-CLIP								
ViT-B/16	*68.95*	*87.62*	*66.69*	*64.49*	*88.13*	*62.08*	*85.45*	*90.88*
ViT-B/32	*63.36*	*86.91*	*64.74*	*62.31*	*83.39*	*58.84*	*81.66*	*90.45*
ViT-L/14	*75.93*	*87.90*	*75.12*	*75.39*	*93.01*	*76.17*	*89.32*	*95.05*

son. We use the existing 8 image classification benchmarks to measure top-1 zero-shot classification accuracy.

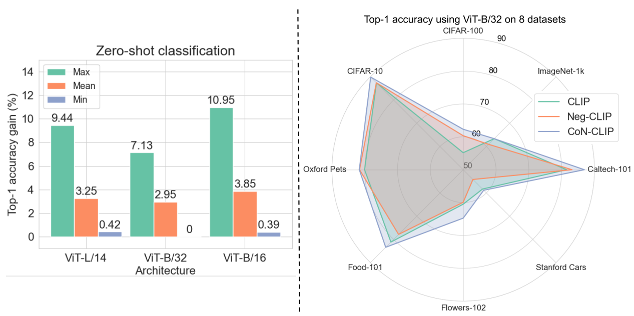

Figure 6. We show performance gains of CoN-CLIP over CLIP across all datasets per architecture (left) and comparisons of image classification using the ViT-B/32 backbone (right).

Results: Table. 4 presents a comprehensive evaluation of all model architectures on all datasets mentioned above. Considering CLIP as baseline, we find that CoN-CLIP shows greater or equal top-1 accuracy for all datasets and architecture. As shown in Fig. 6, CoN-CLIP ViT-B/16 exhibits an average improvement of 3.85% across all datasets, with the highest improvement of 10.95% on the CIFAR-100 dataset. Overall, CoN-CLIP presents an average gain of 3.19% in top-1 accuracy across all datasets and architectures. Additionally, we also use Neg-CLIP as a baseline for the ViT-B/32 architecture and present its performance alongside CoN-CLIP in Fig. 6. We find that Neg-CLIP falls below CLIP ViT-B/32 on top-1 accuracy when evaluated on ImageNet-1k, Stanford Cars, Flowers-102, Food-101. This validates that the fine-tuning process improves CLIP's understanding of negations as well as zero-shot classification.

4.2.3 Compositional Understanding

To understand a scene as a function of its individual components, a model must learn to parse object relations and attributes. Thus, it is necessary to evaluate the framework for negation understanding on data domains specifically designed to benchmark fine-grained compositional understanding. This experiment evaluates the performance of CoN-CLIP on tasks pertaining to attributes and relations in natural scenes. Such an evaluation of generalizability in a different data domain aims to show that learning negations can strengthen overall compositional understanding. We evaluate CoN-CLIP on SugarCREPE and use CLIP as baseline for zero-shot image-text matching. Specifically, a VLM must match a given image to its true caption by correctly rejecting the provided false caption which may contain replaced/added/swapped objects, attributes and relations. Notably, we test CoN-CLIP (trained with CC-Neg) on SugarCREPE without fine-tuning it on any additional data tailored towards compositionality.

Results: As shown in Table. 5, out of 21 total settings, CoN-CLIP outperforms CLIP in 18 settings on SugarCREPE, showing an average improvement of 4.4% in retrieval performance (R@1). In particular, the largest improvements are for the Add fold of SugarCREPE where the average gain in retrieval accuracy is 10.65% for Add-Object, and 9.64% for Add-Attribute. We infer that this occurs due to the implicit effects of the proposed dataset. Considering that negated captions are essentially fine-grained variations of the true captions, learning to repel negated captions in the proposed objective increases the sensitivity of the model to changes in the atoms of input texts. Moreover, it allows CoN-CLIP to pay greater attention to how concepts are composed in text by forcing the model to prioritize associations with correctly composed semantics.

4.2.4 Ablation Study

We conduct an ablation study to evaluate the impact of each loss function, with results summarized in Table 6. Specifically, we report the average R@1 for each fold of SugarCREPE, image-to-caption matching accuracy for the CC-

Table 5. Evaluating CoN-CLIP on SugarCREPE alongside CLIP on R@1. Highest performance for a fold and CLIP backbone are underlined and *italicised* respectively.

Model	**Replace**			**Add**		**Swap**	
	Object	Attribute	Relation	Object	Attribute	Object	Attribute
CLIP							
ViT-B/16	93.28	80.83	*66.00*	78.32	66.61	*59.59*	64.41
ViT-B/32	90.79	80.07	*68.99*	76.91	68.35	60.81	63.06
ViT-L/14	94.06	79.18	65.07	78.17	71.38	60.00	62.16
CoN-CLIP							
ViT-B/16	*93.58*	*80.96*	63.30	*87.29*	*79.62*	59.18	*65.16*
ViT-B/32	*91.76*	*80.96*	66.28	*87.92*	*78.03*	*63.67*	*66.96*
ViT-L/14	*95.31*	*81.72*	*66.99*	*90.15*	*77.60*	*65.36*	*63.06*

Table 6. Our ablation study with the CLIP ViT-B/32 backbone and different combinations of loss terms across all experiments. CC-Neg - $\mathcal{L}_{conclip}$ yields the highest average performance (underlined) across all settings, strongly outperforming CREPE-Negate - $\mathcal{L}_1$.

Dataset - Loss	**SugarCREPE R@1** **Replace**	**Add**	**Swap**	**CC-Neg Accuracy**	**Image classification Top-1 accuracy**
CC-Neg - $\mathcal{L}_1$	79.36	82.22	61.64	99.76	73.32
CC-Neg - $\mathcal{L}_2$	79.38	85.26	65.88	56.07	72.97
CC-Neg - $\mathcal{L}_1 + \mathcal{L}_2$	80.55	83.29	64.07	99.72	73.37
CC-Neg - $\mathcal{L}_{conclip}$	79.67	82.97	65.18	99.70	73.95
CREPE-Negate - $\mathcal{L}_1$	72.40	81.14	61.94	69.79	70.55

Neg dataset, and the average top-1 image classification accuracy across all datasets in Sec. 4.2.2. Additionally, we study the effect of various contrastive loss design choices, and the effect of choosing between CC-Neg and CREPE-Negate data domains. Fine-tuning CLIP on CREPE-Negate with $\mathcal{L}_1$ (using CREPE hard negatives as c'), results in significantly lower performance (refer Table 6). Notably, the effect of $\mathcal{L}_3$ does not seem to stand out in terms of benchmark performance, however, we include distractor images as part of our design towards a more holistic understanding of negation. Since the set of distractor images for even one negated caption can be variable and abstract, $\mathcal{L}_3$ shows little effect on text-driven benchmarks. However, distractor images can still provide useful information towards more general compositionality (see results on the Swap partition of SugarCREPE) and improved image classification.

5. Conclusion

This paper explores the challenges foundational vision-language models (VLMs) face when interpreting negations in textual descriptions. We observe that VLMs frequently overlook negations, leading to incorrect associations between negated text and corresponding images. To address this, we introduce CC-Neg, a novel dataset designed for evaluating negation comprehension, which uses large language models (LLMs) to mine challenging negative text examples. We further present CoN-CLIP, a fine-tuning framework that enhances contrastive learning by incorporating negation-rich captions and distractor images into the training process. Results demonstrate that CoN-CLIP outperforms models like CLIP, Neg-CLIP, FLAVA, and BLIP in recognizing negations. Beyond serving as a benchmark, our work opens new avenues for scalable, data-driven improvements to foundation models, enabling them to better handle underrepresented concepts without the need for vast pre-training datasets. Additionally, CC-Neg can help evaluate and fine-tune generative models, which struggle to comprehend negated text prompts as presented in Fig. 1.

6. Acknowledgement

Vatsa and Singh are partially supported by a research grant from the Ministry of Electronics and Information Technology, Govt. of India.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick,

Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc., 2022. 2

[2] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. 3

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, pages 446–461. Springer, 2014. 2, 6

[4] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23040–23050, 2023. 2

[5] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pretraining to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 2, 3

[6] Keyan Chen, Xiaolong Jiang, Yao Hu, Xu Tang, Yan Gao, Jianqi Chen, and Weidi Xie. Ovarnet: Towards open-vocabulary object attribute recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23518–23527, 2023. 2

[7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020. 2

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 2, 6

[9] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021. 3

[10] Radina Dobreva and Frank Keller. Investigating negation in pre-trained vision-and-language models. In Jasmijn Bastings, Yonatan Belinkov, Emmanuel Dupoux, Mario Giulianelli, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad, editors, *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 350–362, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 2

[11] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178. IEEE, 2004. 2, 6

[12] Matteo Greco. On the syntax of surprise negation sentences: A case study on expletive negation. *Natural Language & Linguistic Theory*, pages 775–825, 2020. 2

[13] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022. 2

[14] Md Mosharaf Hossain, Antonios Anastasopoulos, Eduardo Blanco, and Alexis Palmer. It's not a non-issue: Negation as a source of error in machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 3869–3885, 2020. 2

[15] Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 9106–9118, 2020. 2

[16] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in Neural Information Processing Systems*, 2024. 3, 4

[17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[18] Kenan Jiang, Xuehai He, Ruize Xu, and Xin Eric Wang. Comclip: Training-free compositional image and text matching. *arXiv preprint arXiv:2211.13854*, 2022. 2

[19] Nora Kassner and Hinrich Schütze. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, 2020. 2

[20] Sangeet Khemlani, Isabel Orenes, and Philip N Johnson-Laird. The negations of conjunctions, conditionals, and disjunctions. *Acta Psychologica*, pages 1–7, 2014. 2

[21] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 2, 6

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2, 6

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2, 4

[25] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[26] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 2

[27] 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 *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 3, 5

[28] Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah A Smith, Yejin Choi, and Hannaneh Hajishirzi. Vera: A general-purpose plausibility estimation model for commonsense statements. *arXiv preprint arXiv:2305.03695*, 2023. 3

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[30] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 2019. 2

[31] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921, 2023. 3

[32] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, 2022. 3

[33] M Minderer, A Gritsenko, A Stone, M Neumann, D Weissenborn, A Dosovitskiy, A Mahendran, A Arnab, M Dehghani, Z Shen, et al. Simple open-vocabulary object detection with vision transformers. arxiv 2022. *arXiv preprint arXiv:2205.06230*. 2

[34] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pages 529–544. Springer, 2022. 2

[35] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. *Advances in Neural Information Processing Systems*, pages 13988–14000, 2021. 2

[36] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 2, 6

[37] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[38] Isabel Orenes. "looking at" negation: Faster processing for symbolic rather than iconic representations. *Journal of Psycholinguistic Research*, (6):1417–1436, 2021. 2

[39] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3498–3505. IEEE, 2012. 2, 6

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 2019. 6

[41] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 4, 5

[42] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. *arXiv preprint arXiv:2204.06125*, 2022. 2

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[44] Amir Rosenfeld, Richard S. Zemel, and John K. Tsotsos. The elephant in the room. *CoRR*, 2018. 2

[45] Tara Safavi, Jing Zhu, and Danai Koutra. Negater: Unsupervised discovery of negatives in commonsense knowledge bases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5633–5646, 2021. 2

[46] Claudia Schon, Sophie Siebert, and Frieder Stolzenburg. Negation in cognitive reasoning. In *German Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 217–232. Springer, 2021. 2

[47] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2

[48] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. *CoRR*, 2021. 2, 4

[49] Alberto Testoni, Claudio Greco, and Raffaella Bernardi. Artificial intelligence models do not ground negation, humans do. guesswhat?! dialogues as a case study. *Frontiers in big Data*, page 736709, 2022. 2, 5

[50] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, (2):64–73, 2016. 3

[51] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 3

[52] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1802–1812, 2023. 3

[53] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 5696–5710. Curran Associates, Inc., 2022. 2

[54] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022. 3

[55] 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 *International Conference on Learning Representations*, 2022. 2, 3, 4

[56] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 2, 6