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

This supplementary material elaborates on our contributions and methodology and presents additional results. First, we provide details regarding our data generation process in Sec. A. Then, Sec. B explains the methodology of obtaining distractor images for fine-tuning CLIP on our novel objectives. Lastly, Sec. C provides additional results related to our experiments: negation understanding, zero-shot image classification, and general compositional understanding (replacing, adding, and swapping object, attributes, and relations).

A. Prompting Scheme for CC-Neg Generation

We use the image-labels subset of CC-3M as source for our image-caption pairs. Following this, the construction of CC-Neg is divided into two parts - negated caption generation and negative image mining.

A.1. Generation of negated captions in CC-Neg

The positive prompts from CC-3M are first decomposed using LLMs for relation parsing, and subsequently converted to negated captions, following the prompts and the process depicted in Fig. 1. To summarize:

1. Positive image-caption pairs are extracted from the image-labels split of CC-3M.
2. First, we prompt PaLM-2 to parse the relations inside positive prompts, i.e., captions related to the images. Each prompt is decomposed into a subject - who/what the sentence is about, as well as multiple predicate-object pair. Each predicate qualifies actions, or relates an object to the subject to specify a state of being.
3. A single predicate-object pair is randomly selected to be negated for each sample.
4. We then employ PaLM-2 to replace the selected predicate with one of *no, not* and *without* appropriately, and combine the subsequent atoms and relations into a negative caption.
5. We drop samples with greater than 9 predicate-object pairs as this level of complexity is rarely found in real world data and is irrelevant from a human standpoint.

B. Distractor Images for Finetuning

This section describes our process of obtaining distractor images used in our fine-tuning process. Given a true caption c and a negated caption c' from CC-Neg, we first segregate concepts we know to be present from concepts that are absent in the scene, depicted in c'. Specifically, we take the subject s and the negated object o_n from the relation parsing output of PaLM-2 while generating c'. Next, given s to be present and o_n to be absent in the scene depicted by c', we select a distractor image I' from a large set of images (MSCOCO [3]) $\boldsymbol{X}$ by

$$I' = \underset{x \in \boldsymbol{X}}{\arg\max} \quad \phi(x, s) - \phi(x, o_n) \tag{1}$$

where $\phi(\cdot, \cdot)$ is the CLIP similarity function. Inspecting the results of this process, we obtain distractor images which represent the subject well but do not represent the negated object and its predicate. Examples of these distractor images are show in Fig. 2. While the negated captions can have more interpretations than their corresponding distractor images, we find distractor images to be suitable image negatives aimed at disentangling the existing effect of negations as well as other compositional deficiencies. Notably, we do not consider this method of mining distractor images as a viable substitute for our framework CoN-CLIP. This is because (i) it requires an initial decomposition step to identify which semantics are present or absent, and (ii) it cannot be used to learn the effect of negations for a VLM. Consequently, this method sacrifices speed and acquisition of important new knowledge. Further, it cannot be used effectively for improving negation understanding in downstream applications of VLMs such as multimodal large language models (MLLMs) and text-to-image generation models (as mentioned in the main manuscript). This method is simply used to find suitable examples in the image modality which anchor the embeddings of negated captions for improved semantic disentanglement (with $\mathcal{L}_1$ and $\mathcal{L}_2$).

C. Additional Results

This section presents additional results with an added baseline: LaCLIP [1], a variant of CLIP which adds text augmentations during pretraining. This leads to a primary ben-

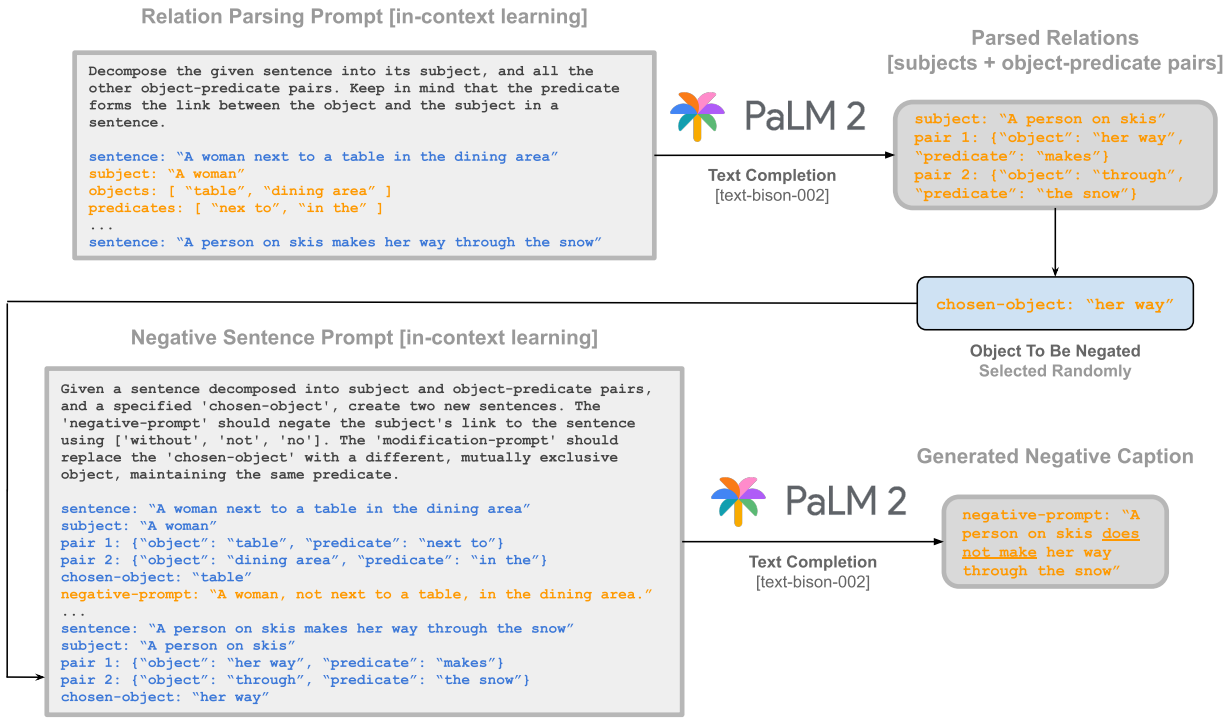

Figure 1. Illustration of using in-context learning with PaLM-2 to parse relations and subsequently generate negated captions. One meta-input-output demonstration is shown colored in blue and orange, each indicating the input and the output, respectively. Decomposition of positive captions into a subject and multiple predicate-object pairs is shown at the top. A random predicate-object pair is then selected and negated to generate a negative prompt, shown at the bottom.

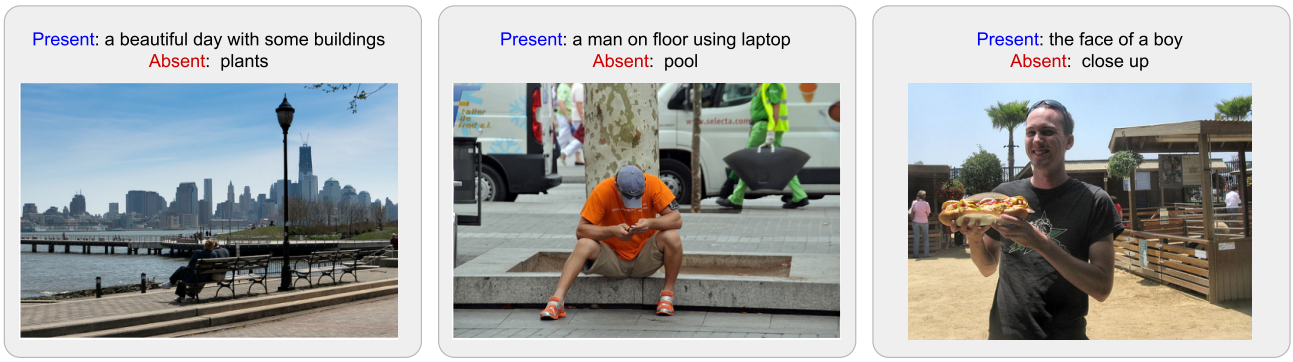

Figure 2. Examples of distractor images obtained from our process are shown above.

efit to image classification by reducing overfitting to specific prompts [1]. Due to more exposure to language formats, we expect LaCLIP to show improved text understanding and compositionality, however, this is not an application proposed in [1]. Hence, to avoid confusion and maintain focus on the core contributions of CoN-CLIP, we omit this baseline from the main manuscript and provide the same here for comprehensiveness. We show LaCLIP's performance on negation understanding, general purpose compositionality, and zero-shot image classification. These results are shown alongside CLIP [4], NegCLIP [6], BLIP [2], FLAVA [5], and all variants of CoN-CLIP, *i.e.*, $\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_{12}$, and $\mathcal{L}_{conclip}$.

Model configuration	No. of predicate-object pairs				
	#1 (166)	#2 (160)	#3 (75)	#4 (28)	#5 (8)
CLIP	70.48	68.13	69.33	53.57	37.5
LaCLIP	69.88	81.25	64.0	67.86	50.0
BLIP	76.51	63.12	72.0	42.86	50.0
FLAVA	62.05	58.75	57.33	57.14	62.5
NegCLIP	69.28	64.38	64.0	57.14	25.0
CoN-CLIP	96.99	93.13	92.0	82.14	75.0

Table 1. Model accuracies over CC-Neg evaluation subset samples that use "no" to specify negation. Scores are calculated over subset splits with the same number of predicate-object pairs, indicated by the number over each column. Split sizes are denoted inside parentheses.

Model Configuration	No. of predicate-object pairs				
	#1 (11910)	#2 (10081)	#3 (3047)	#4 (722)	#5 (188)
CLIP	68.13	64.94	59.11	59.83	56.91
LaCLIP	71.08	66.52	58.62	54.02	53.19
BLIP	60.76	60.07	54.25	56.23	50.00
FLAVA	59.73	54.58	50.51	50.42	52.13
Neg-CLIP	66.32	63.21	57.66	55.12	55.32
CoN-CLIP	99.88	99.84	99.64	99.03	99.47

Table 2. Model accuracies over CC-Neg evaluation subset samples that use "not" to specify negation. Scores are calculated over subset splits with the same number of predicate-object pairs, indicated by the number over each column. Split sizes are denoted inside parentheses.

Model Configuration	No. of predicate-object pairs				
	#1 (4850)	#2 (5466)	#3 (2328)	#4 (680)	#5 (203)
CLIP	66.14	66.39	65.38	63.09	70.94
LaCLIP	69.36	59.79	56.36	55.44	48.28
BLIP	70.12	67.54	64.18	62.06	67.98
FLAVA	65.98	63.57	61.77	61.62	58.13
Neg-CLIP	62.12	60.92	57.39	56.32	60.59
CoN-CLIP	99.88	99.82	99.91	99.71	100.0

Table 3. Model accuracies over CC-Neg evaluation subset samples that use "without" to specify negation. Scores are calculated over subset splits with the same number of predicate-object pairs, indicated by the number over each column. Split sizes are denoted in parentheses.

C.1. Negation Understanding

We present thorough evaluation of all VLMs on CC-Neg and its attributes: number of predicate-object pairs $\mathcal{K}$ and type of negation word used. These results are given in Table 1, Table 2, Table 3. Additionally, the various settings of CoN-CLIP are also evaluated on the same, the results of which are given in Table 4, Table 5, and Table 6.

C.2. Compositionality with SugarCREPE

We add LaCLIP as a baseline for compositional understanding on the SugarCREPE benchmark in Table 7. All CoN-CLIP settings are also evaluated on SugarCREPE in Table 8.

C.3. Zero-shot Image Classification

We use LaCLIP as a baseline in zero-shot image classification in Table 9.

C.4. Fine-tuned Baseline Evaluation

We fine-tune CLIP on only true-pairings (CLIP-FT) and evaluate this model alongside the proposed CoN-CLIP approach as another baseline.

Model Configuration	No. of predicate-object pairs				
	#1 (166)	#2 (160)	#3 (75)	#4 (28)	#5 (8)
CoNCLIP ViT-B/32 $\mathcal{L}_1$	96.99	93.13	90.67	85.71	62.5
CoNCLIP ViT-B/32 $\mathcal{L}_2$	65.06	58.75	57.33	32.14	25.0
CoNCLIP ViT-B/32 $\mathcal{L}_{12}$	95.18	93.13	90.67	78.57	75.0
CoN-CLIP ViT-B/32 $\mathcal{L}_{conclip}$	96.99	93.13	92.0	82.14	75.0

Table 4. CoN-CLIP ablation study over CC-Neg evaluation subset samples that use "no" to specify negation. Scores are calculated over subset splits with the same number of predicate-object pairs, indicated by the number over each column.

Model Configuration	No. of predicate-object pairs				
	#1 (11910)	#2 (10081)	#3 (3047)	#4 (722)	#5 (188)
CoNCLIP ViT-B/32 $\mathcal{L}_1$	99.84	99.84	99.74	99.58	99.47
CoNCLIP ViT-B/32 $\mathcal{L}_2$	55.89	52.16	49.56	48.20	46.81
CoNCLIP ViT-B/32 $\mathcal{L}_{12}$	99.83	99.81	99.67	99.17	98.94
CoN-CLIP ViT-B/32 $\mathcal{L}_{conclip}$	99.88	99.84	99.64	99.03	99.47

Table 5. CoN-CLIP ablation study over CC-Neg evaluation subset samples that use "not" to specify negation. Scores are calculated over subset splits with the same number of predicate-object pairs, indicated by the number over each column.

Model Configuration	No. of predicate-object pairs				
	#1 (4850)	#2 (5466)	#3 (2328)	#4 (680)	#5 (203)
CoNCLIP ViT-B/32 $\mathcal{L}_1$	99.94	99.89	99.79	99.85	100.0
CoNCLIP ViT-B/32 $\mathcal{L}_2$	62.93	61.23	58.76	55.00	60.59
CoNCLIP ViT-B/32 $\mathcal{L}_{12}$	99.92	99.87	99.74	99.85	100.0
CoN-CLIP ViT-B/32 $\mathcal{L}_{conclip}$	99.88	99.82	99.91	99.71	100.0

Table 6. CoN-CLIP ablation study over CC-Neg evaluation subset samples that use "without" to specify negation. Scores are calculated over subset splits with the same number of predicate-object pairs, indicated by the number over each column.

References

[1] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[2] 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

[3] 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. 1

[4] 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

[5] 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

[6] 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

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
LaCLIP							
ViT-B/16	93.22	79.69	58.32	77.44	66.18	59.59	59.15
ViT-B/32	91.28	77.66	57.75	75.41	64.01	55.51	59.55
ViT-L/14	93.28	81.09	61.73	81.57	73.55	62.44	58.70
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 7. 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
CoN-CLIP							
$\mathcal{L}_1$	91.71	80.58	65.79	84.82	79.62	59.18	64.11
$\mathcal{L}_2$	91.65	81.22	65.29	88.02	82.51	64.49	67.27
$\mathcal{L}_{12}$	91.46	81.85	68.35	86.81	79.77	61.63	66.52
$\mathcal{L}_{conclip}$	91.76	80.96	66.28	87.92	78.03	63.67	66.96

Table 8. Evaluating CoN-CLIP ablations on SugarCREPE.

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*
LaCLIP								
ViT-B/16	67.20	87.39	66.11	*67.82*	82.82	*85.61*	83.95	91.82
ViT-B/32	62.01	*87.95*	62.76	*65.56*	75.61	*80.01*	80.84	*91.38*
ViT-L/14	74.50	*89.81*	*75.85*	*79.32*	90.28	*90.77*	89.29	*96.69*
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

Table 9. Evaluation of CoN-CLIP ($\mathcal{L}_{conclip}$) 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	Mean Δ (%) $\uparrow$
CLIP	0.98
LaCLIP	-0.99
NegCLIP	-1.01
CoN-CLIP	62.03

Table 10. Comparing Δ values averages across 8 image classification datasets (Sec. 5.1) of the main manuscript using ViT-B/32.

Model	ImageNet 1k	Caltech 101	Flowers 102	CIFAR 100	Food 101	Stanford Cars	Oxford Pets	CIFAR 10
CLIP	63.36	81.50	60.50	55.18	81.15	58.33	80.08	88.97
CLIP-FT	63.60	85.92	64.14	61.45	83.76	58.92	81.28	89.27
CoN-CLIP	63.36	86.91	64.74	62.31	83.39	58.84	81.66	90.45

Table 11. Zero-shot Image Classification (B-32): CoN-CLIP outperforms CLIP-FT on 5 out of 8 datasets and is comparable on the rest.

Model	Replace			Add		Swap		CC-Neg
	Object	Attribute	Relation	Object	Attribute	Object	Attribute	
CLIP	90.79	80.07	68.99	76.91	68.35	60.81	63.06	65.70
CLIP-FT	91.71	81.59	64.29	85.93	79.91	60.40	66.06	61.47
CoN-CLIP	91.89	82.74	66.57	85.21	79.62	61.63	66.21	99.70

Table 12. Evaluation on SugarCREPE and CC-Neg (B-32). CoN-CLIP retains its preformance gain against CLIP-FT on SugarCREPE.