WACV 2025
TUCSON, ARIZONA • FEB 28 - MAR 4

PROJECT PAGE

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, [2]Metafusion, [3]IIT Jodhpur, [4]Lehigh University

Motivation

- Negations are necessary to specify absence of concepts.
- VLMs such as CLIP ignore negation words: "no", "not", "without".

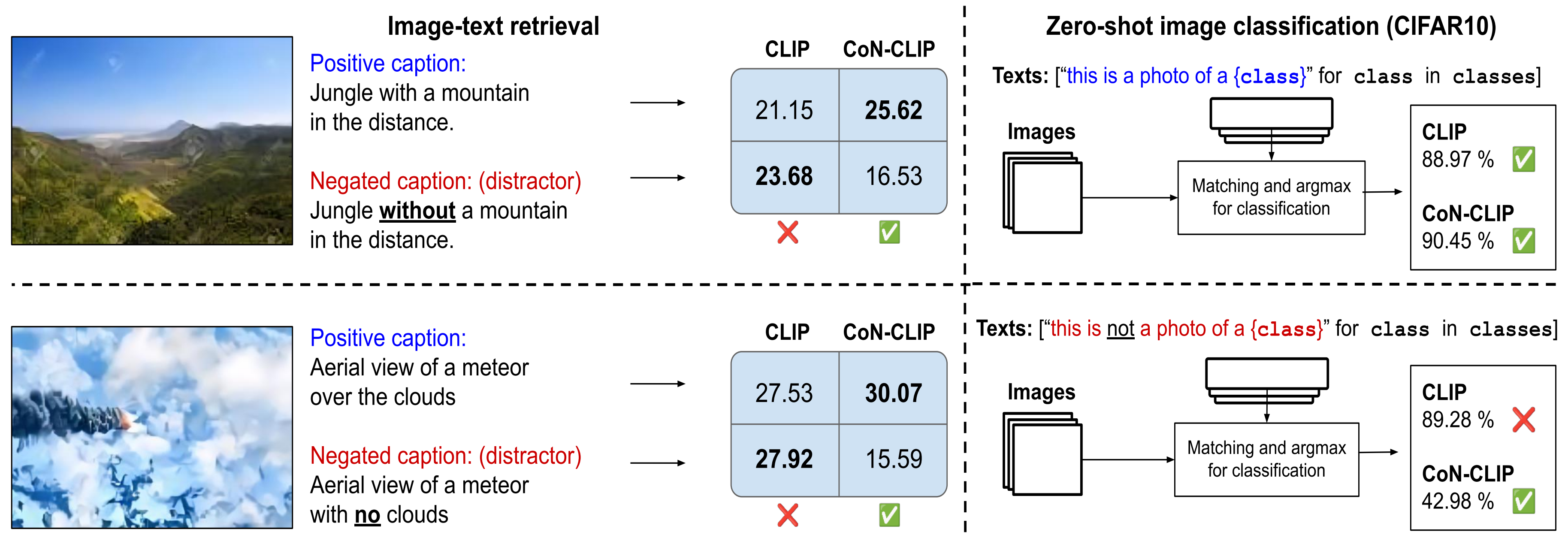


Image-text retrieval

	CLIP	CoN-CLIP
Positive caption: Jungle with a mountain in the distance.	21.15	**25.62**
Negated caption: (distractor) Jungle **without** a mountain in the distance.	**23.68**	16.53
	✗	✓
Positive caption: Aerial view of a meteor over the clouds	27.53	**30.07**
Negated caption: (distractor) Aerial view of a meteor with **no** clouds	**27.92**	15.59
	✗	✓

Zero-shot image classification (CIFAR10)

Texts: [“this is a photo of a `{class}`” for `class` in `classes`]

Images → Matching and argmax for classification → CLIP 88.97 % ✓ CoN-CLIP 90.45 % ✓

Texts: [“this is not a photo of a `{class}`” for `class` in `classes`]

Images → Matching and argmax for classification → CLIP 89.28 % ✗ CoN-CLIP 42.98 % ✓

- Contrastive objective rewards bag-of-words behavior.
- Negations are rarely represented in pretraining data.
- Poor negation understanding affects downstream applications of VLMs: T2I generation, I2T generation.

Text-to-Image Generative Models

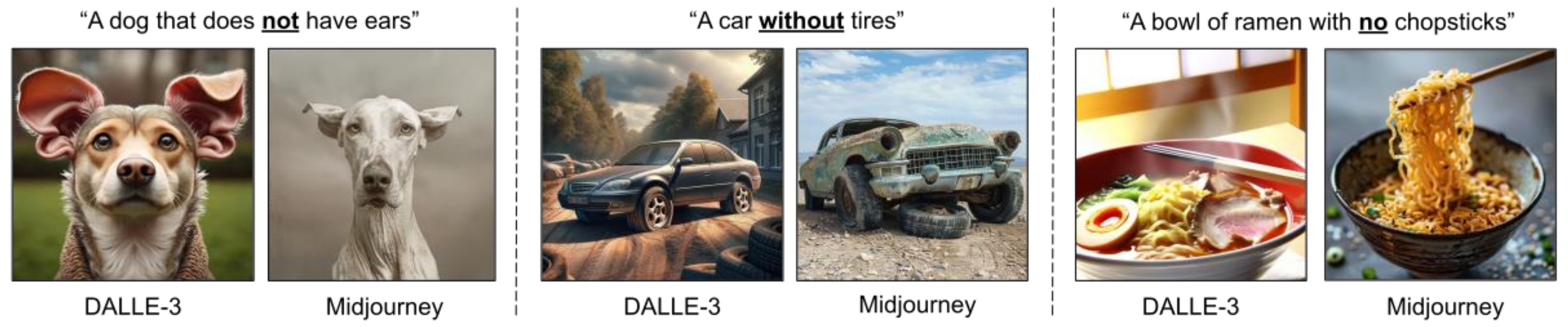
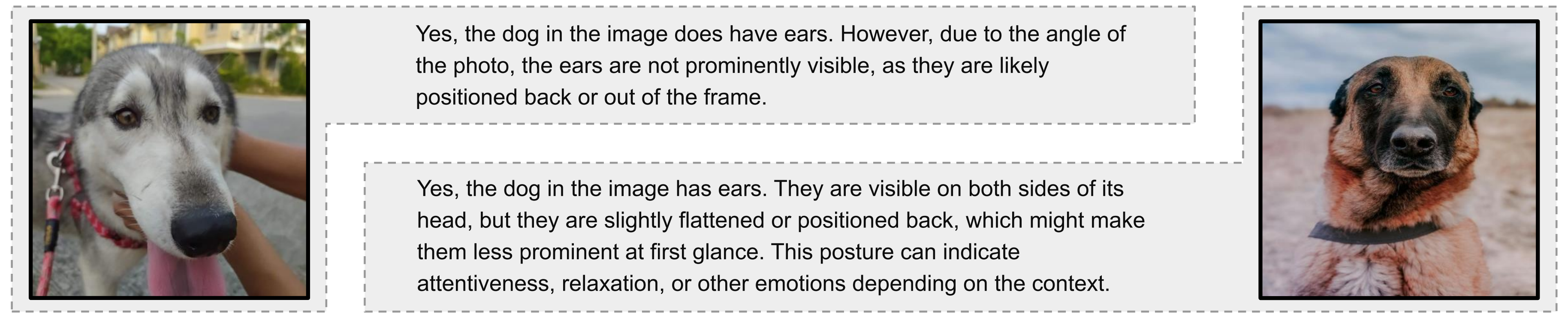


Image-to-text Generative Models

Prompt: “Does this dog have ears?”

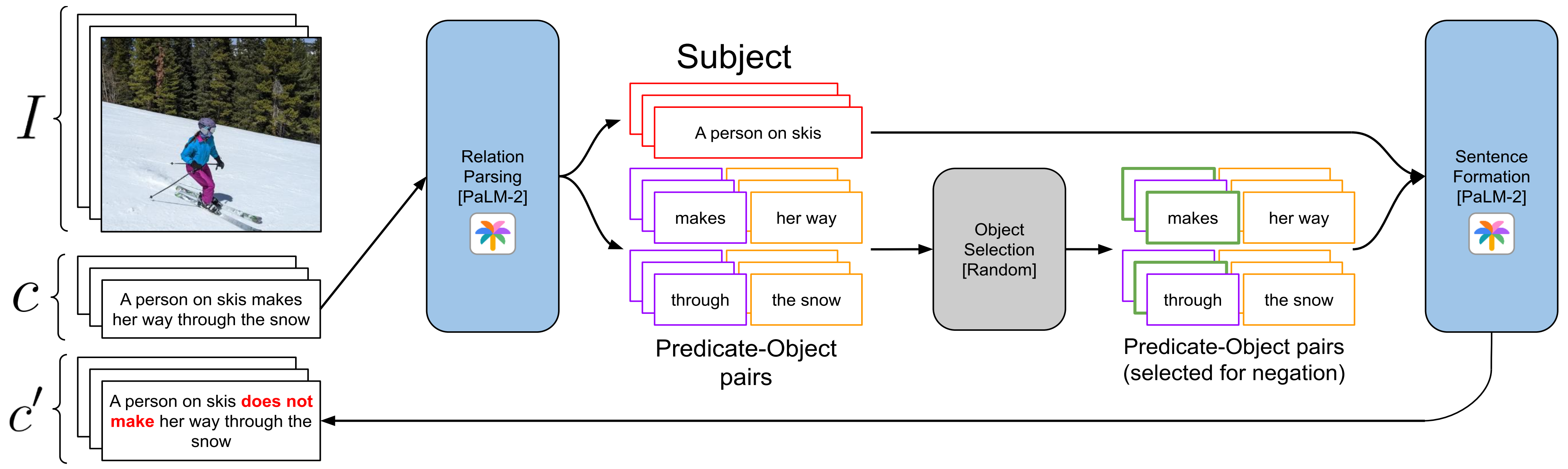


Yes, the dog in the image does have ears. However, due to the angle of the photo, the ears are not prominently visible, as they are likely positioned back or out of the frame.

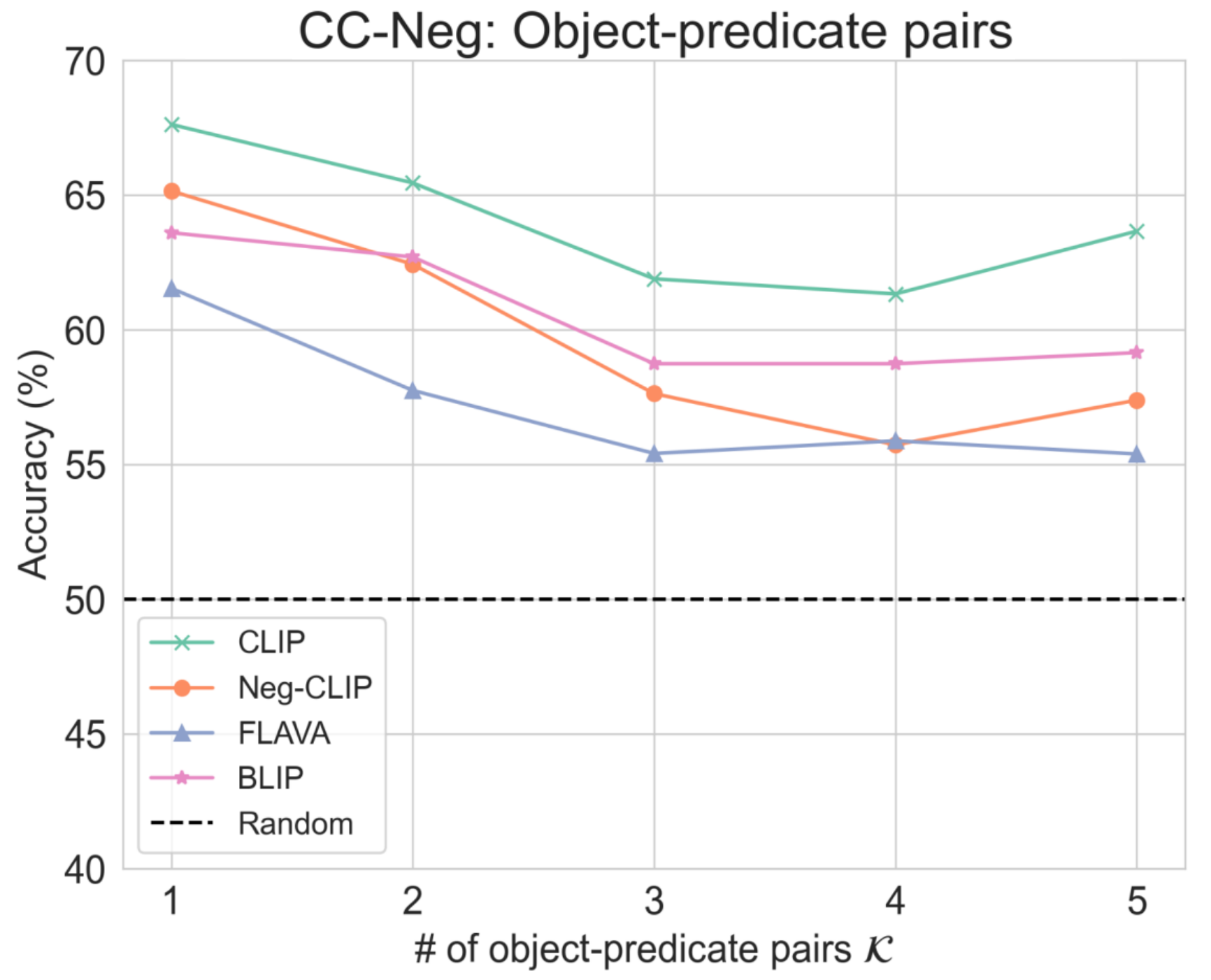
Yes, the dog in the image has ears. They are visible on both sides of its head, but they are slightly flattened or positioned back, which might make them less prominent at first glance. This posture can indicate attentiveness, relaxation, or other emotions depending on the context.

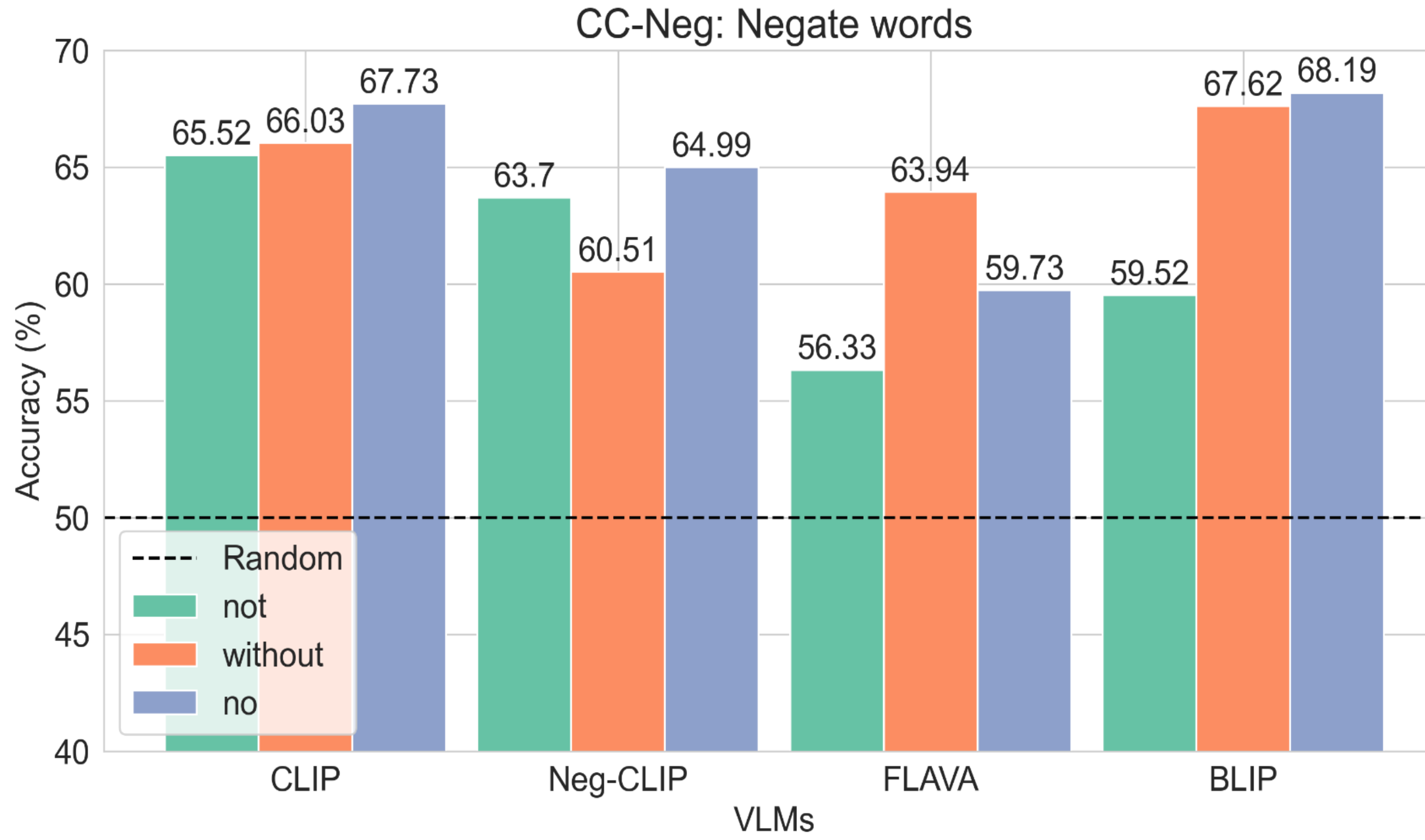
CC-Neg: Benchmark for Negations

- CC-Neg: a large-scale high-quality dataset containing negations for image-text matching.
    - Parse the caption of an image and negate a concept (negated caption)
    - Subset of CC-3M using with PaLM-2 to write negated captions by subject-object-predicate decomposition of true caption.
    - Yields 228,246 (Image, True Caption, Negated Caption) triplets


I
c
A person on skis makes her way through the snow
c'
A person on skis does not make her way through the snow
Relation Parsing [PaLM-2]
Subject
A person on skis
makes
her way
through
the snow
Predicate-Object pairs
Object Selection [Random]
makes
her way
through
the snow
Predicate-Object pairs (selected for negation)
Sentence Formation [PaLM-2]

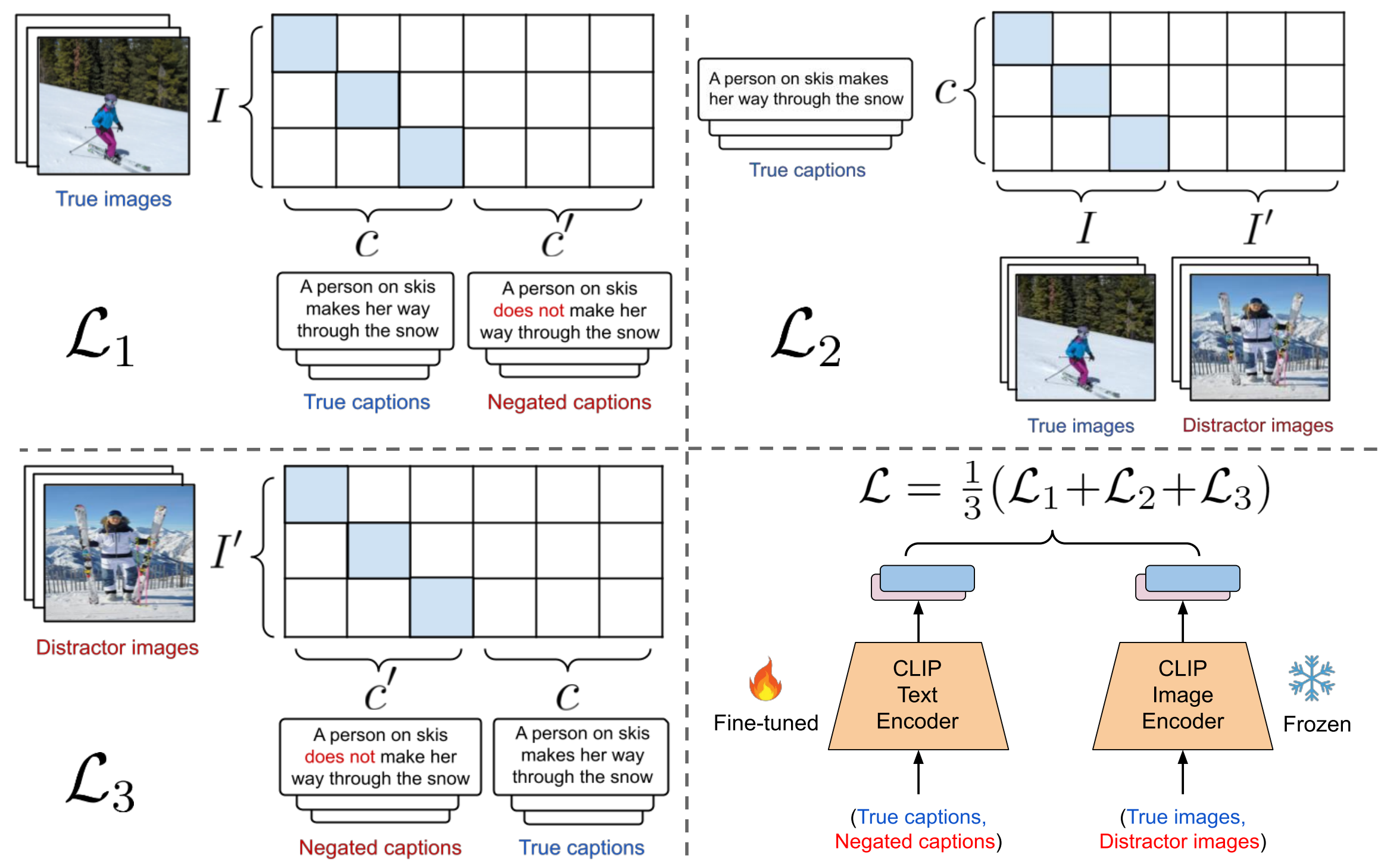
- CC-Neg benchmarks various VLMs on negation understanding and shows:
    - VLMs fail to recognize negations
    - Negation understanding degrades with negated caption complexity
    - VLMs favor certain negation words (e.g. "no") over others (e.g. "without")





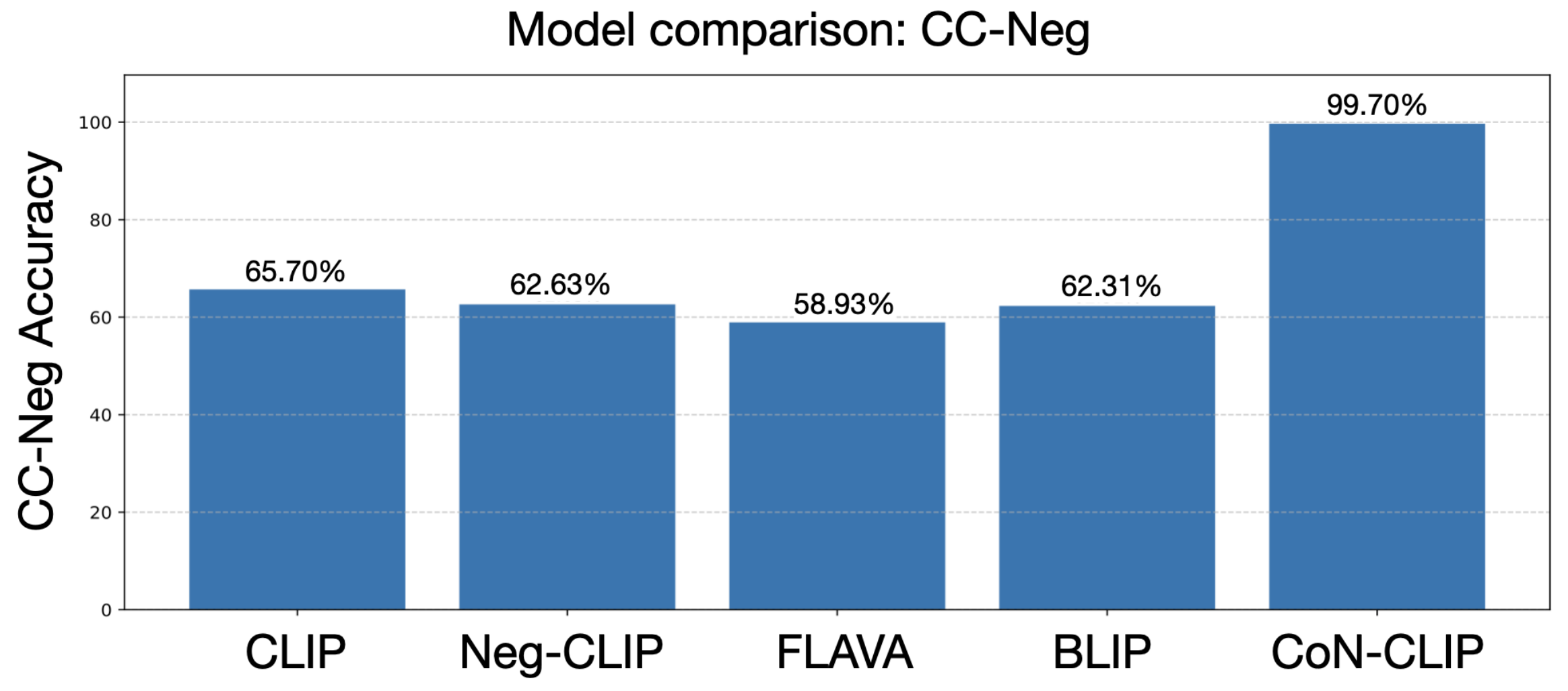
CoN-CLIP: Fine-tuning CLIP on Negations

- CC-Neg can be used to impart negation understanding.
- CLIP fine-tuned with custom objective using
    - Image + true caption
    - Negated caption
    - Mined distractor images as reflection of negated caption


True images
I
c
c'
A person on skis makes her way through the snow
A person on skis does not make her way through the snow
True captions
Negated captions
$\mathcal{L}_1$
A person on skis makes her way through the snow
True captions
c
I
I'
True images
Distractor images
$\mathcal{L}_2$
Distractor images
I'
c'
c
A person on skis does not make her way through the snow
A person on skis makes her way through the snow
Negated captions
True captions
$\mathcal{L}_3$
Fine-tuned
CLIP Text Encoder
(True captions, Negated captions)
CLIP Image Encoder
Frozen
(True images, Distractor images)

$$\mathcal{L} = \frac{1}{3}(\mathcal{L}_1+\mathcal{L}_2+\mathcal{L}_3)$$

- CoN-CLIP shows strong understanding of negations on held-out CC-Neg eval set:



- CoN-CLIP shows improved performance on popular VLM tasks:

SugarCREPE image-text matching

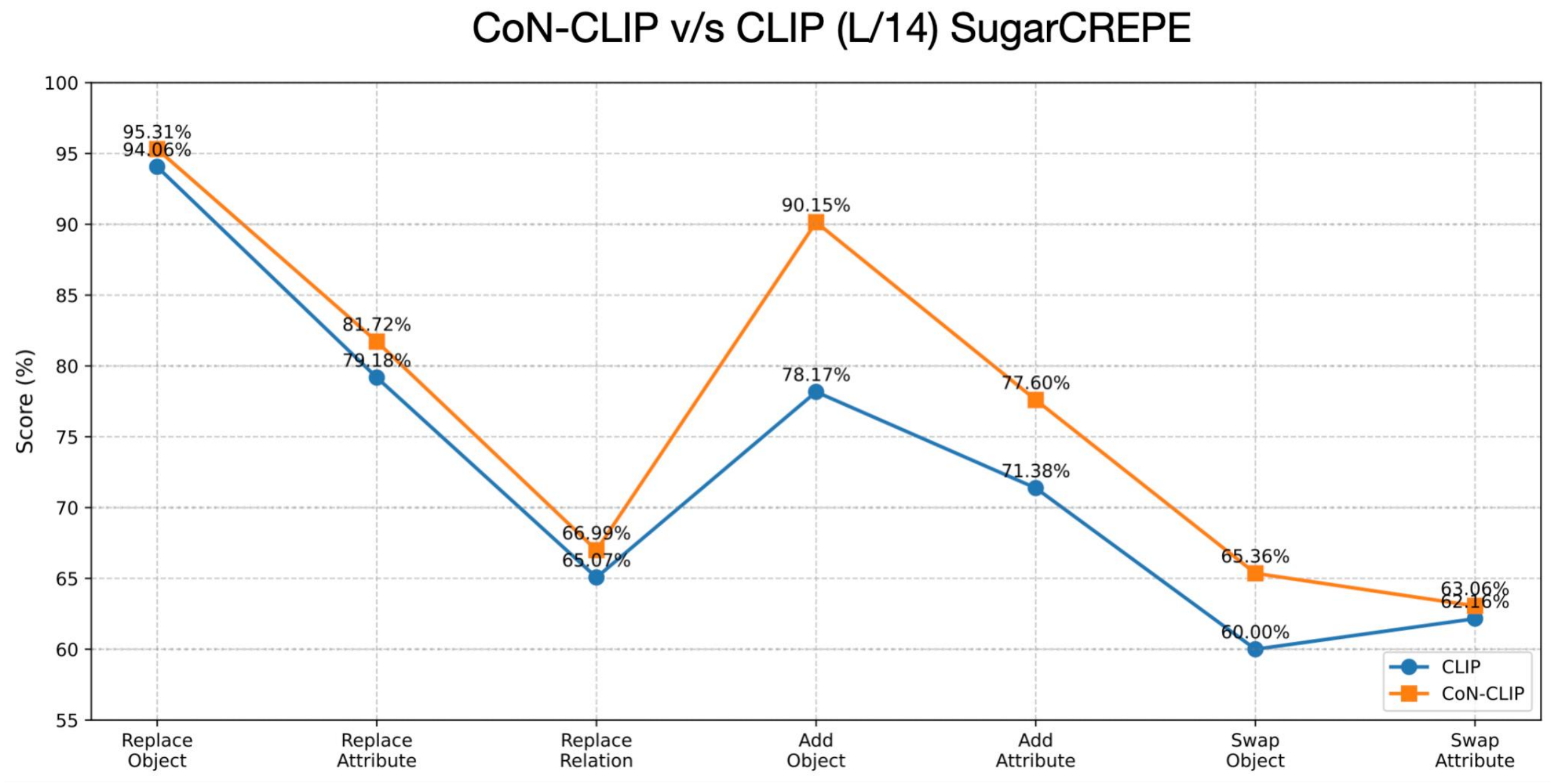
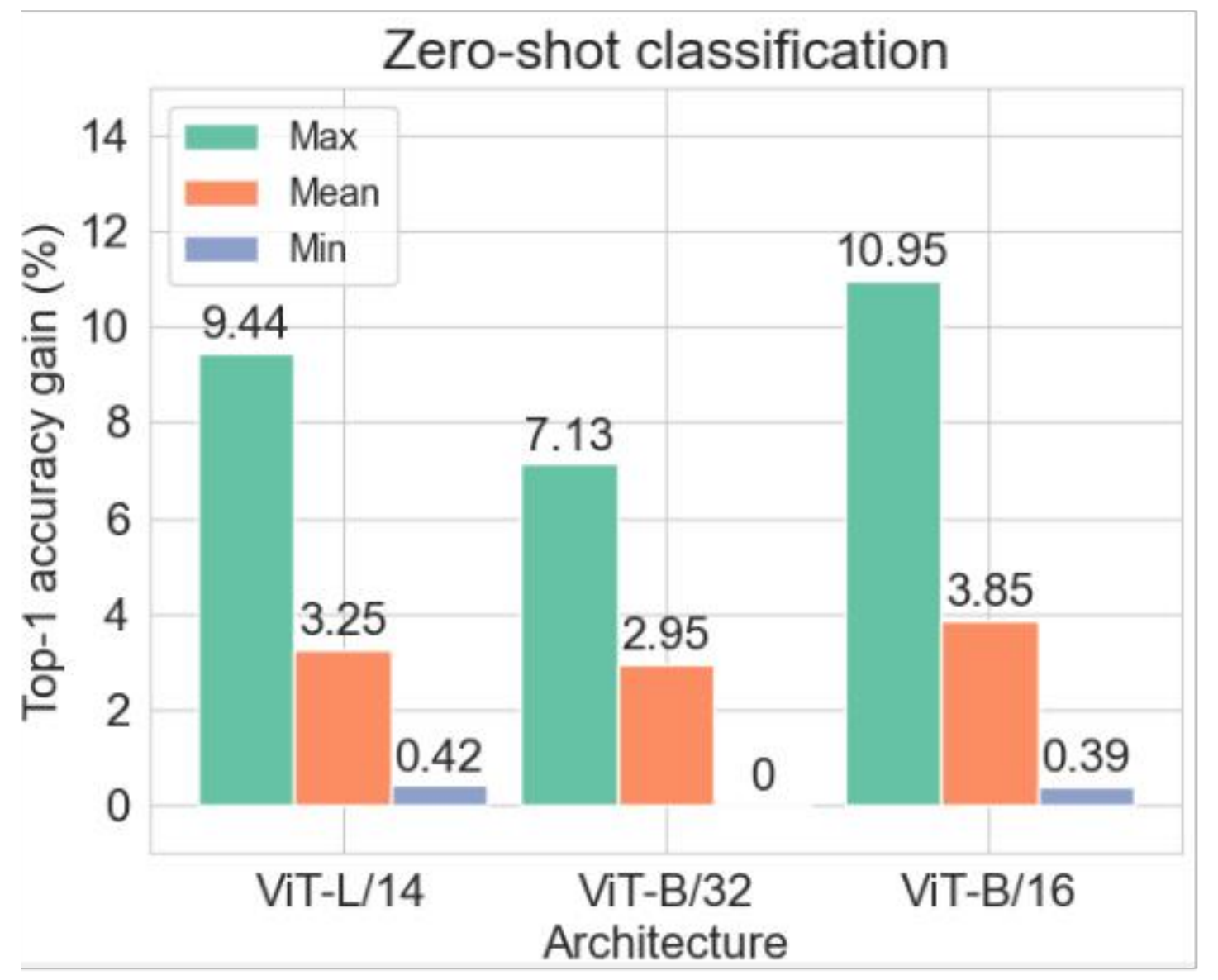


Image classification (8 datasets)


Zero-shot classification
Top-1 accuracy gain (%)
Max
Mean
Min
ViT-L/14
ViT-B/32
ViT-B/16
Architecture