Know “No” Better:
A Data-Driven Approach for Enhancing Negation Awareness in CLIP

Supplementary Material

A. OpenCLIP Training Dataset Statistics

In this section, we provide additional statistics on negation-related terms in the datasets used for OpenCLIP [1] training, complementing the analysis presented in Sec. 2.2. Specifically, we report the frequency of the negation terms “no,” “not,” and “without” in the DataComp-1B [3] and LAION-2B [13] datasets in Tab. 1. Both datasets exhibit trends similar to those observed in LAION-400M [13]. These findings are consistent with the statistics we reported for LAION-400M, highlighting the insufficient representation of negation in datasets used for OpenCLIP pre-training.

Table 1. Proportion of negation in captions and words within DataComp-1B and LAION-2B.

Dataset	Level	Total Count	Neg. Count	Neg. Ratio
DataComp-1B [3]	Caption	1.38B	10.4M	0.75%
	Word	13.8B	11.8M	0.09%
LAION-2B [13]	Caption	2.08B	19.3M	0.93%
	Word	21.9B	21.1M	0.10%

B. More Ablation Studies

In addition to the results presented in Sec. 4.3, we further evaluate the impact of using original captions instead of the generated captions for fine-tuning on the same set of images.

The results shown in Tab. 2 demonstrate that incorporating our generated captions consistently achieves the highest performance on both the VALSE and NegRefCOCOg benchmarks, across all architectures. Notably, while fine-tuning with original captions can also lead to degradation, as observed in the ViT-B/32 architecture on the VALSE benchmark, fine-tuning with our generated captions consistently improves performance. This underscores the efficacy of our data generation pipeline in enhancing negation comprehension.

Table 2. More ablation results on different data configurations.

Arch.	Data Config.	VALSE ↑	NegRefCOCOg ↑
ViT-B/32	Original	70.97	57.73
	+ Original Caption	68.35	60.45
	+ Our Caption	**80.15**	**64.09**
ViT-B/16	Original	69.48	58.64
	+ Original Caption	73.97	60.91
	+ Our Caption	**80.52**	**64.32**
ViT-L/14	Original	66.85	57.27
	+ Original Caption	74.91	60.00
	+ Our Caption	**79.59**	**62.95**
ViT-L/14 @336px	Original	64.61	57.05
	+ Original Caption	73.97	58.41
	+ Our Caption	**80.34**	**62.95**

C. Data Generation Pipelines

We provide additional details on the two data generation pipelines discussed in Sec. 3.1 and Sec. 3.2 of the main paper. Specifically, we include the prompts used with the LLM and MLLM during data generation and provide examples of the generated image-caption pairs. The prompts used in the two pipelines are presented in Tab. 3 and Tab. 4 respectively.

We provide qualitative examples of the data generated by our proposed data generation pipelines in Fig. 1. As shown in Fig. 1 (a), captions generated by Pipeline 1 accurately incorporate the absence of objects such as “car,” “ball,” or “curtains,” which are contextually plausible within the scene. This process enhances the training data with negation terms while maintaining alignment with the image content. Fig. 1 (b) demonstrates how Pipeline 2 captures a broader scope of negation, such as negating actions (e.g., “not swimming”), adjectival phrases (e.g., “not in the wild”). These examples highlight the flexibility and effectiveness of this pipeline in generating rich negation-inclusive captions.

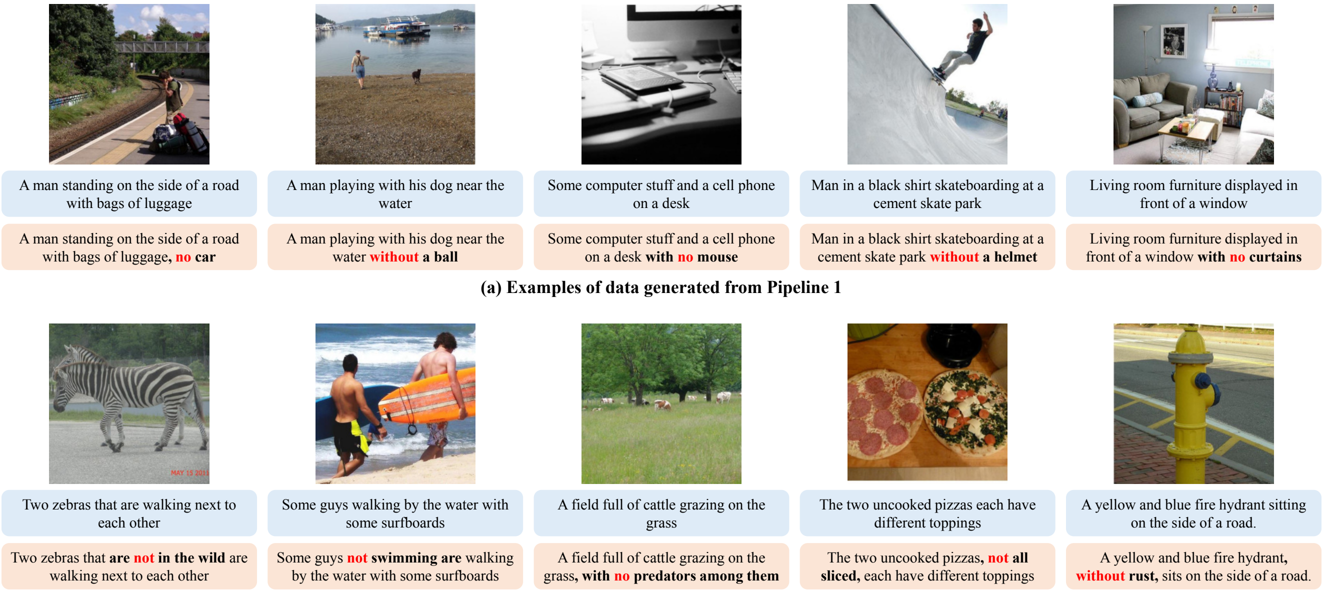
D. NegRefCOCOg Benchmark

In this section, we elaborate details on the construction of our proposed NegRefCOCOg benchmark in Sec. 3.3.

Selection Criteria To construct NegRefCOCOg, we begin by selecting samples from the RefCOCOg [17] dataset that meet the following criteria:

- The original image patch P_o^+, corresponding to the negation-inclusive prompt T, has a height and width of at least 100 pixels.
- At least one other image patch belonging to the same category as P_o^+ has a height and width of at least 100 pixels and does not overlap with P_o^+. We then designate one of these patches as P_o^-.

Image Patch Maximization To ensure alignment with T, we maximize the sizes of P_o^+ and P_o^- under the following constraints. As a result, we obtain the final patches, P^+ and


A man standing on the side of a road with bags of luggage
A man standing on the side of a road with bags of luggage, no car
A man playing with his dog near the water
A man playing with his dog near the water without a ball
Some computer stuff and a cell phone on a desk
Some computer stuff and a cell phone on a desk with no mouse
Man in a black shirt skateboarding at a cement skate park
Man in a black shirt skateboarding at a cement skate park without a helmet
Living room furniture displayed in front of a window
Living room furniture displayed in front of a window with no curtains
Two zebras that are walking next to each other
Two zebras that are not in the wild are walking next to each other
Some guys walking by the water with some surfboards
Some guys not swimming are walking by the water with some surfboards
A field full of cattle grazing on the grass
A field full of cattle grazing on the grass, with no predators among them
The two uncooked pizzas each have different toppings
The two uncooked pizzas, not all sliced, each have different toppings
A yellow and blue fire hydrant sitting on the side of a road.
A yellow and blue fire hydrant, without rust, sits on the side of a road.
MAY 18 2014

(a) Examples of data generated from Pipeline 1

(b) Examples of data generated from Pipeline 2

Figure 1. Examples of data generated by our proposed pipelines. (a) demonstrates captions augmented through Pipeline 1, and (b) illustrates captions augmented through Pipeline 2. **Blue boxes** represent the original captions, and **red boxes** show the augmented captions with negation terms.

Table 3. Prompts used in our Pipeline 1.

Step 1: Extracting Plausible Object from Caption	
System	`You are a helpful chatbot that answers with only one word.`
User	`Name an object that is not mentioned in the caption, but is likely to be in the image corresponding to the caption '{caption}'.`
LLM	`{object}.`
Step 2: Verifying Object Absence with MLLM	
System	`A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions.`

User	`<image>`
	`Is there {object} in this image? Answer either yes or no.`

MLLM	`{yes/no}.`
Step 3: Augmenting Caption with Negation	
System	`You are a helpful chatbot that generates concise caption.`
User	`Add the absence of the {object} to the caption '{caption}'.`
LLM	`{updated_caption}.`

Table 4. Prompts used in our Pipeline 2.

Step 2: Augmenting Caption with Negation	
System	`You are a helpful chatbot that generates concise caption.`
User	`When the answer to the question {question} is 'no', reconstruct the caption '{caption}'.`
LLM	`{updated_caption}.`

P^-, where P^+ is the expanded version of P_o^+ and P^- is the expanded version of P^-.

- The expanded patch must not overlap with the other patch before its maximization, which can be expressed as:

$$P^+ \cap P_o^- = \varnothing, \quad P^- \cap P_o^+ = \varnothing.$$

- Horizontal expansion is limited to the original width of the patch in each direction (left and right).
- Vertical expansion is limited to the original height of the patch in each direction (top and bottom).

We obtain 440 triplets of (T, P^+, P^-) through this process, where P^+ is well-aligned with T and P^- serves as challenging hard negative for evaluation. These 440 samples form the NegRefCOCOg benchmark, which we use to evaluate the ability of models to handle negation comprehensively and accurately.

Fig. 2 illustrates examples from NegRefCOCOg. These examples demonstrate the diversity of negation scenarios in NegRefCOCOg, including object absence, action negation with various negation terms.

E. Results on Additional Models

Our method is model-agnostic and applies to SigLIP [18], showing consistent gains on negation benchmarks while preserving general performance, as shown in Tab. 5.

Table 5. Comparison of model performance on negation and general benchmarks across different architectures of SigLIP [18].

Model	**Arch.**	**Negation Benchmarks**		**General Benchmarks**	
		VALSE	**NegRefCOCOg**	**ImageNet**	**COCO**
SigLIP	base	67.98	61.59	**75.69**	72.10
Neg.SigLIP		**77.15**	**68.18**	75.10	**77.30**
SigLIP	large	73.03	62.05	79.73	75.20
Neg.SigLIP		**79.40**	**67.73**	**79.76**	**80.10**
SigLIP	so400m	75.47	63.18	**82.26**	76.30
Neg.SigLIP		**84.27**	**67.05**	81.68	**81.70**

F. Expanded Discussion on T2I Generation

F.1. Methods

In T2I generation, research has emerged focusing on the removal of unwanted concepts from generated images. [2, 5, 16]. These methods typically preprocess text prompts by identifying and removing negation-related components, reformulating the prompt to exclude negation. For example, given the text prompt "a panda in a forest without flowers," LMD [5] uses an LLM is used to identify "flowers" as the negated object and remove it, resulting in a final layout-based prompt that retains only "a panda" and "a forest."

In contrast, our approach directly uses the original prompt "a panda in a forest without flowers" without any preprocessing, allowing the T2I model to process negation as part of the input text. This provides a different perspective on handling negation, where the model learns to interpret negation within natural language rather than relying on preprocessing to modify the prompt.

F.2. Evaluation

Existing methods evaluate negation comprehension in T2I models using the LMD [5] Negation benchmark, which consists of 10 prompts structured as "a realistic photo of a scene without [object]" and uses object detectors to verify object absence.

We provide evaluation results of NegationCLIP on the LMD Negation benchmark in Tab. 6. Standard SD struggles significantly with negation, achieving only 20% accuracy, indicating that it often fails to remove the negated object. In contrast, simply replacing SD's text encoder with NegationCLIP's text encoder improves accuracy to 94%, successfully removing the negated object in nearly all cases. Unsurprisingly, applying LMD [5] method to SD yields perfect performance, as it heuristically removes the negation-related part from the final prompt, ensuring that the object does not appear in the generated image.

Table 6. Results on LMD Negation benchmark.

Method	**LMD Negation**
SD	20%
SD w/ NegationCLIP text encoder	94%
SD + LMD [5]	100%

However, while the LMD Negation benchmark effectively evaluates object removal, it does not account for negation beyond object presence or absence, such as actions and attributes. While it can determine whether an object has been successfully removed, it fails to capture cases where negation modifies an entity's state (e.g., *a dog not running*) or its properties (e.g., *a not blue sphere*). To address this limitation, we introduce more diverse prompts that encompass negation in objects, actions, and attributes as shown in Tab. 8).

Since object detectors are insufficient for evaluating these more complex negation cases, we drew inspiration from the TIFA [4] metric's MLLM-based VQA system and introduce the Neg Score as a more comprehensive evaluation measure in Sec 5.2.

G. Experiment Details

G.1. CelebA Classification

In Tab. 7, we provide attribute-specific prompts for all 40 attributes of CelebA [7] in our experiment in Sec. 2.1. NegationCLIP improves the accuracy from 60.8% to 64.0%, whereas CoN-CLIP [14] sees a decline to 60.6%. We note that achieving significantly higher accuracy on CelebA is inherently challenging, as the dataset includes subjective attributes (e.g., *attractive*).

G.2. Fine-Tuning Configuration

We detail the fine-tuning configurations used to train our CLIP [11] models with the data generated from the proposed data generation pipeline.

The generated dataset was split into 80% for training and 20% for validation. We used batch sizes of 512 for ViT-B/32, 256 for ViT-B/16, 128 for both ViT-L/14 and ViT-L@336px, and 64 for ViT-BigG/14 which we fine-tuned specifically for the text-to-image (T2I) generation experiment using SDXL-1.0 [10] in Sec. 5.1. All models were

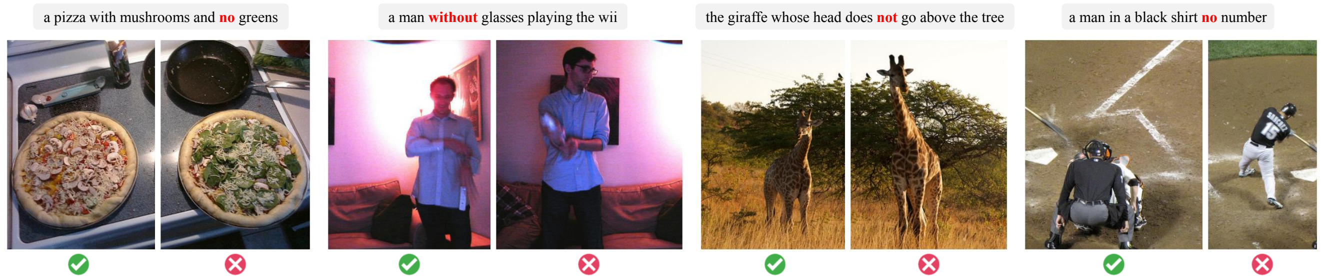

a pizza with mushrooms and no greens
a man without glasses playing the wii
the giraffe whose head does not go above the tree
a man in a black shirt no number

Figure 2. Examples from the proposed NegRefCOCOg benchmark. Each example consists of a textual description containing negation (top) and two image patches: the correct patch aligned with the negated description (green checkmark) and the incorrect patch representing a challenging hard negative (red cross).

fine-tuned using a single NVIDIA L40 GPU.

G.3. T2I Generation

We provide details on T2I experiments shown in Sec. 5.1.

For generating negation-inclusive prompts, we utilized ChatGPT [9] to construct 107 prompts. To evaluate whether the generated images accurately reflected the given prompts, we employed ChatGPT to create two corresponding questions for each prompt. The first question was designed to assess whether the subject of the prompt was correctly generated, with the expected answer being “yes.” The second question aimed to verify whether the negation-related aspect was properly removed, with the expected answer being “no.” For example, for the prompt “a man not wearing a hat,” the two generated questions were “Is this a man?” and “Is the man wearing a hat?”. The complete list of prompts and their corresponding questions is provided in Tab. 8.

In SD-1.4 [12], we replaced its CLIP ViT-L/14 text encoder with CoN-CLIP text encoder and our fine-tuned negation-aware text encoder of the same architecture. For SDXL, which employs two text encoders (CLIP ViT-L/14 and CLIP ViT-BigG/14), we substituted both encoders with our fine-tuned versions.

G.4. Referring Image Segmentation

For our experiments in referring image segmentation in Sec. 5.2, we utilized the publicly available weight of CLIPSeg [8] trained on the PhraseCut [15] dataset as our baseline. Without any additional training, we replaced the text encoder in the CLIPSeg architecture with CoN-CLIP ViT-B/16 text encoder and our fine-tuned CLIP ViT-B/16 text encoder.

We followed CLIPSeg to determine the threshold values for binary segmentation, setting it to 0.3 for experiments on the PhraseCut dataset and 0.1 for experiments on Ref-COCOg (Neg), which is based on the COCO [6] dataset.

H. Additional Qualitative Results

Fig. 3 demonstrates additional qualitative examples from T2I task. For each prompt, we include all images generated using 5 different random seeds.

For both SDXL [10] and SD-1.4 [12], substituting the original CLIP text encoder with NegationCLIP text encoder improves the ability to accurately reflect negation in generated images. Notably, SD-1.4, which uses only one text encoder, maintains high-quality image generation despite the substitution, highlighting that the fine-tuned NegationCLIP text encoder preserves overall image quality while enhancing negation comprehension.

Certain challenging prompts, such as “a city without buildings” or “a museum without exhibits,” expose limitations in the model’s ability to remove concepts with strong biases (e.g., buildings in city contexts, exhibits in museum contexts). These limitations may stem from the limited representation of such cases in the training data for generative models. Identifying the underlying causes of these limitations and addressing them constitutes an important direction for future work.

Table 7. CelebA attribute-specific prompts and balanced accuracy

Attribute	**Positive Prompt**	**Negative Prompt**
5 o Clock Shadow	a photo of a person with a 5 o'clock shadow	a photo of a person with no 5 o'clock shadow
Arched Eyebrows	a photo of a person with arched eyebrows	a photo of a person with not arched eyebrows
Attractive	a photo of an attractive person	a photo of a not attractive person
Bags Under Eyes	a photo of a person with bags under eyes	a photo of a person with no bags under eyes
Bald	a photo of a bald person	a photo of a not bald person
Bangs	a photo of a person with bangs	a photo of a person with no bangs
Big Lips	a photo of a person with big lips	a photo of a person with not big lips
Big Nose	a photo of a person with a big nose	a photo of a person with a not big nose
Black Hair	a photo of a person with black hair	a photo of a person with not black hair
Blond Hair	a photo of a person with blond hair	a photo of a person with not blond hair
Blurry	a blurry photo of a person	a not blurry photo of a person
Brown Hair	a photo of a person with brown hair	a photo of a person with not brown hair
Bushy Eyebrows	a photo of a person with bushy eyebrows	a photo of a person with not bushy eyebrows
Chubby	a photo of a chubby person	a photo of a not chubby person
Double Chin	a photo of a person with a double chin	a photo of a person with no double chin
Eyeglasses	a photo of a person wearing glasses	a photo of a person not wearing glasses
Goatee	a photo of a person with goatee	a photo of a person with no goatee
Gray Hair	a photo of a person with gray hair	a photo of a person with not gray hair
Heavy Makeup	a photo of a person with heavy makeup	a photo of a person with no heavy makeup
High Cheekbones	a photo of a person with high cheekbones	a photo of a person with not high cheekbones
Male	a photo of a male	a photo of a not male
Mouth Slightly Open	a photo of a person with mouth slightly open	a photo of a person with mouth not slightly open
Mustache	a photo of a person with mustache	a photo of a person with no mustache
Narrow Eyes	a photo of a person with narrow eyes	a photo of a person with not narrow eyes
No Beard	a photo of a person with no beard	a photo of a person with beard
Oval Face	a photo of a person with oval face	a photo of a person with not oval face
Pale Skin	a photo of a person with pale skin	a photo of a person with not pale skin
Pointy Nose	a photo of a person with a pointy nose	a photo of a person with not a pointy nose
Receding Hairline	a photo of a person with a receding hairline	a photo of a person with no receding hairline
Rosy Cheeks	a photo of a person with rosy cheeks	a photo of a person with not rosy cheeks
Sideburns	a photo of a person with sideburns	a photo of a person with no sideburns
Smiling	a photo of a person smiling	a photo of a person not smiling
Straight Hair	a photo of a person with straight hair	a photo of a person with not straight hair
Wavy Hair	a photo of a person with wavy hair	a photo of a person with not wavy hair
Wearing Earrings	a photo of a person wearing earrings	a photo of a person not wearing earrings
Wearing Hat	a photo of a person wearing a hat	a photo of a person not wearing a hat
Wearing Lipstick	a photo of a person wearing lipstick	a photo of a person not wearing lipstick
Wearing Necklace	a photo of a person wearing a necklace	a photo of a person not wearing a necklace
Wearing Necktie	a photo of a person wearing a necktie	a photo of a person not wearing a necktie
Young	a photo of a young person	a photo of a not young person

Table 8. Text-to-image generation prompts and corresponding questions.

Prompt	Question 1	Question 2
a man not wearing a hat	Is this a man?	Is the man wearing a hat?
a dog not running	Is this a dog?	Is the dog running?
a not white cat	Is this a cat?	Is the cat white?
a room with no window	Is this a room?	Is there a window in the room?
a cat with no whiskers	Is this a cat?	Does the cat have whiskers?
a park with no benches	Is this a park?	Are there benches in the park?
a woman without glasses	Is this a woman?	Is the woman wearing glasses?
a car not parked in the driveway	Is this a car?	Is the car parked in the driveway?
a man with no hat	Is this a man?	Is the man wearing a hat?
a garden with no flowers	Is this a garden?	Are there flowers in the garden?
a lake with no boats	Is this a lake?	Are there boats on the lake?
a tree not in bloom	Is this a tree?	Is the tree in bloom?
a room without furniture	Is this a room?	Is there furniture in the room?
a street with no people	Is this a street?	Are there people on the street?
a cup not filled with coffee	Is this a cup?	Is the cup filled with coffee?
a phone not on the table	Is this a phone?	Is the phone on the table?
a man not wearing shoes	Is this a man?	Is the man wearing shoes?
a city skyline with no skyscrapers	Is this a city skyline?	Are there skyscrapers in the city skyline?
a woman not smiling	Is this a woman?	Is the woman smiling?
a car without wheels	Is this a car?	Does the car have wheels?
a road with no signs	Is this a road?	Are there signs on the road?
a bridge without railings	Is this a bridge?	Does the bridge have railings?
a sky without clouds	Is this a sky?	Are there clouds in the sky?
a playground with no swings	Is this a playground?	Are there swings in the playground?
a building without windows	Is this a building?	Does the building have windows?
a shop with no customers	Is this a shop?	Are there customers in the shop?
a bike not leaning against a wall	Is this a bike?	Is the bike leaning against a wall?
a train station with no trains	Is this a train station?	Are there trains at the train station?
a shelf with no books	Is this a shelf?	Are there books on the shelf?
a person with no backpack	Is this a person?	Does the person have a backpack?
a room not filled with light	Is this a room?	Is the room filled with light?
a school without students	Is this a school?	Are there students in the school?
a cat without a tail	Is this a cat?	Does the cat have a tail?
a forest with no leaves	Is this a forest?	Are there leaves in the forest?
a chair not facing the table	Is this a chair?	Is the chair facing the table?
a beach without sand	Is this a beach?	Is there sand on the beach?
a wall with no decorations	Is this a wall?	Are there decorations on the wall?
a man not reading a book	Is this a man?	Is the man reading a book?
a street with no lights	Is this a street?	Are there lights on the street?
a riverbank with no trees	Is this a riverbank?	Are there trees on the riverbank?
a boat with no sails	Is this a boat?	Does the boat have sails?
a stadium with no players	Is this a stadium?	Are there players in the stadium?
a store without shelves	Is this a store?	Are there shelves in the store?
a city without buildings	Is this a city?	Are there buildings in the city?
a road without any turns	Is this a road?	Are there turns on the road?
a dog not fetching a ball	Is this a dog?	Is the dog fetching a ball?
a car with no passengers	Is this a car?	Are there passengers in the car?
a child not drinking milk	Is this a child?	Is the child drinking milk?
a tree with no leaves	Is this a tree?	Does the tree have leaves?
a train without passengers	Is this a train?	Are there passengers on the train?
a photo with no people	Is this a photo?	Are there people in the photo?
a river not flowing	Is this a river?	Is the river flowing?
a path without any footprints	Is this a path?	Are there footprints on the path?
a building with no entrance	Is this a building?	Is there an entrance to the building?
a woman not wearing a mask	Is this a woman?	Is the woman wearing a mask?
a bird not flying	Is this a bird?	Is the bird flying?
a not blue sphere	Is this a sphere?	Is the sphere blue?
a dog without a collar	Is this a dog?	Does the dog have a collar?
a child not holding a toy	Is this a child?	Is the child holding a toy?
a bird not perched on a branch	Is this a bird?	Is the bird perched on a branch?
a table with no chairs around	Is this a table?	Are there chairs around the table?
a beach without any umbrellas	Is this a beach?	Are there umbrellas on the beach?
a road not crowded with cars	Is this a road?	Is the road crowded with cars?
a person not holding an umbrella	Is this a person?	Is the person holding an umbrella?
a house without a roof	Is this a house?	Does the house have a roof?
a mountain with no snow	Is this a mountain?	Is there snow on the mountain?
a dog not barking	Is this a dog?	Is the dog barking?
a kitchen without any food	Is this a kitchen?	Is there food in the kitchen?
a forest with no animals	Is this a forest?	Are there animals in the forest?
a desk without a computer	Is this a desk?	Is there a computer on the desk?
a restaurant with no tables	Is this a restaurant?	Are there tables in the restaurant?
a field without crops	Is this a field?	Are there crops in the field?
a living room with no couch	Is this a living room?	Is there a couch in the living room?
a stadium with no spectators	Is this a stadium?	Are there spectators in the stadium?
a child not wearing shoes	Is this a child?	Is the child wearing shoes?
a river with no fish	Is this a river?	Are there fish in the river?
a cup with no handle	Is this a cup?	Does the cup have a handle?
a man not wearing a tie	Is this a man?	Is the man wearing a tie?
a book with no cover	Is this a book?	Does the book have a cover?
a garden without any trees	Is this a garden?	Are there trees in the garden?
a stage with no performers	Is this a stage?	Are there performers on the stage?
a museum without exhibits	Is this a museum?	Are there exhibits in the museum?
a restaurant not serving food	Is this a restaurant?	Is the restaurant serving food?
a market without any vendors	Is this a market?	Are there vendors in the market?
a path with no signs	Is this a path?	Are there signs on the path?
a car with no headlights	Is this a car?	Does the car have headlights?
a person not holding a bag	Is this a person?	Is the person holding a bag?
a house with no doors	Is this a house?	Does the house have doors?
a bird not singing	Is this a bird?	Is the bird singing?
a dog not playing fetch	Is this a dog?	Is the dog playing fetch?
a sidewalk with no pedestrians	Is this a sidewalk?	Are there pedestrians on the sidewalk?
a classroom with no desks	Is this a classroom?	Are there desks in the classroom?
a yard without grass	Is this a yard?	Is there grass in the yard?
a cat not purring	Is this a cat?	Is the cat purring?
a woman not holding a purse	Is this a woman?	Is the woman holding a purse?
a sky with no stars	Is this a sky?	Are there stars in the sky?
a man not holding a briefcase	Is this a man?	Is the man holding a briefcase?
a painting without colors	Is this a painting?	Does the painting have colors?
a lawn with no flowers	Is this a lawn?	Are there flowers on the lawn?
a bridge without any lights	Is this a bridge?	Are there lights on the bridge?
a garden with no vegetables	Is this a garden?	Are there vegetables in the garden?
a person without a shadow	Is this a person?	Does the person have a shadow?
a bus stop without a bench	Is this a bus stop?	Is there a bench at the bus stop?
a cafe with no tables	Is this a cafe?	Are there tables in the cafe?
a street without any trees	Is this a street?	Are there trees on the street?
a mountain with no trails	Is this a mountain?	Are there trails on the mountain?
a field with no animals	Is this a field?	Are there animals in the field?

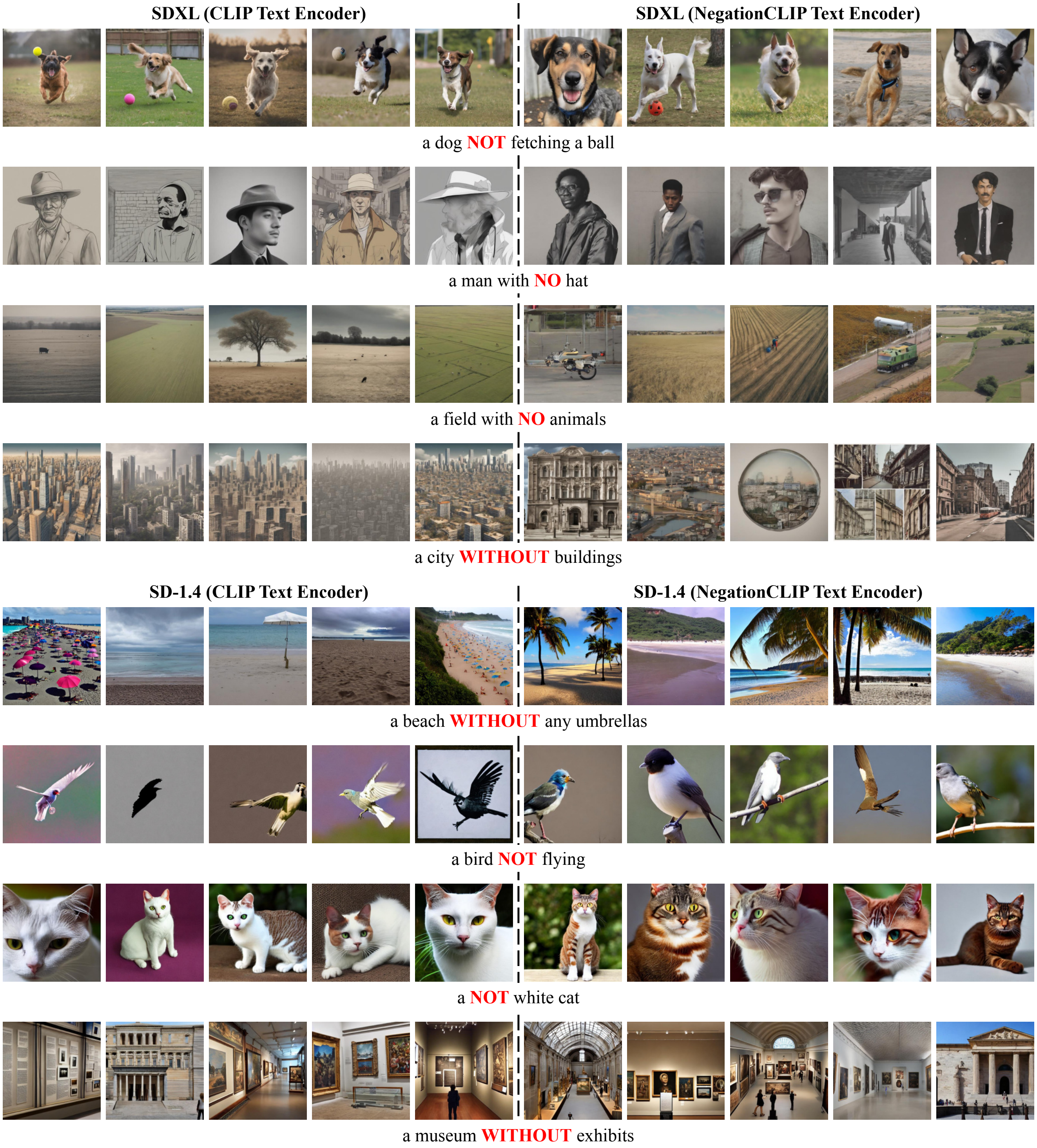

SDXL (CLIP Text Encoder)
SDXL (NegationCLIP Text Encoder)
a dog NOT fetching a ball
a man with NO hat
a field with NO animals
a city WITHOUT buildings
SD-1.4 (CLIP Text Encoder)
SD-1.4 (NegationCLIP Text Encoder)
a beach WITHOUT any umbrellas
a bird NOT flying
a NOT white cat
a museum WITHOUT exhibits

Figure 3. Additional examples for text-to-image generation tasks using Stable Diffusion XL (SDXL) [10] and Stable Diffusion 1.4 (SD-1.4) [12]. Comparisons are shown between models using the original CLIP text encoder and the fine-tuned NegationCLIP text encoder.

References

[1] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 1

[2] Peiran Dong, Song Guo, Junxiao Wang, Bingjie Wang, Jiewei Zhang, and Ziming Liu. Towards test-time refusals via concept negation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[3] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[4] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023. 3

[5] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023. 3

[6] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 4

[7] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 3

[8] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7086–7096, 2022. 4

[9] OpenAI. Gpt-4 technical report, 2023. 4

[10] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3, 4, 7

[11] 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. 3

[12] 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 (CVPR)*, pages 10684–10695, 2022. 4, 7

[13] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 1

[14] Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. Learn" no" to say" yes" better: Improving vision-language models via negations. *arXiv preprint arXiv:2403.20312*, 2024. 3

[15] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225, 2020. 4

[16] Tsung-Han Wu, Long Lian, Joseph E Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting llm-controlled diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6327–6336, 2024. 3

[17] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 1

[18] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 3