TNG-CLIP:
Training-Time Negation Data Generation for Negation Awareness of CLIP

Anonymous ACL submission

Abstract

Vision-language models (VLMs), such as CLIP, have demonstrated strong performance across a range of downstream tasks. However, CLIP is still limited in negation understanding: the ability to recognize the absence or exclusion of a concept. Existing methods address the problem by using a large language model (LLM) to generate large-scale data of image captions containing negation for further fine-tuning CLIP. However, these methods are both time- and compute-intensive, and their evaluations are typically restricted to image-text matching tasks. To expand the horizon, we (1) introduce a training-time negation data generation pipeline such that negation captions are generated during the training stage, which only increases 2.5% extra training time, and (2) we propose the first benchmark, NEG-TTOI, for evaluating text-to-image generation models on prompts containing negation, assessing model's ability to produce semantically accurate images. We show that our proposed method, *TNG-CLIP*, achieves SOTA performance on diverse negation benchmarks of image-to-text matching, text-to-image retrieval, and image generation.

1 Introduction

Vision-language models (VLM), such as CLIP (Radford et al., 2021), provide an efficient approach to tackle vision-language tasks by learning the features of different modalities in a shared embedding space. However, these models fundamentally lack a robust understanding of **negation**—the ability to recognize the absence or exclusion of a concept, *e.g., "A dog **not** playing a ball.", "There is **no** tree on the street."*. Negation is a fundamental aspect of human reasoning, enabling precise description of constraints and expectations in communication. Without proper negation understanding, VLMs generate and retrieve semantically incorrect content, particularly in complicated scenarios where the presence or absence of specific elements critically alters meanings.

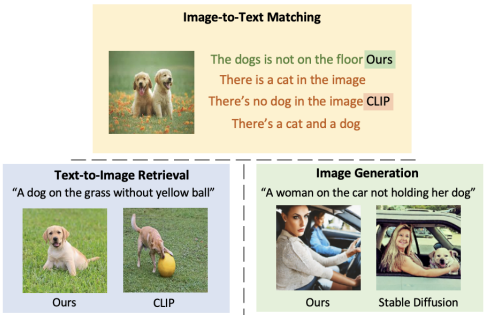

Figure 1: We present *TNG-CLIP*, a negation-aware CLIP that achieves outstanding negation understanding in image-to-text matching, text-to-image retrieval and proposed image generation NEG-TTOI benchmarks.

To tackle this problem, current methods (Alhamoud et al., 2025a; Singh et al., 2024; Park et al., 2025; Yuksekgonul et al., 2023) focus on generating well-designed image-text datasets, such that there are negation captions associated with each image sample, and then fine-tune the underlying VLM. However, such approaches face three challenges: (1) the negation of each caption is designed, generated, and verified via LLMs. Considering the fact that the existing vision-language datasets (Chen et al., 2015; Changpinyo et al., 2021) contain millions of samples, generating the negation dataset is extremely time- and compute-consuming. (2) Unlike standard semantic descriptions, which are typically grounded in observable features, the negation process introduces arbitrariness by specifying the absence of concepts that are not depicted. For example, given an image of *"a dog playing a ball"*, one could construct multiple valid negation captions such as *"a dog playing a ball while no man is present"* or *"a dog playing a ball but not on the beach"*. By generating fixed negation captions, previous methods may constrain the diversity of negation scenarios, thus harming the generalization

of the fine-tuned VLM on negation understanding tasks. (3) Previous methods are mainly evaluated on image-to-text matching and text-to-image retrieval tasks. Considering the versatility of CLIP, however, evaluation should not be constrained to matching-based tasks and must include more diverse downstream tasks such as generation-based tasks, where the text encoder can be used as part of a generative model (Rombach et al., 2022).

We propose a new data generation and training pipeline which generates negation captions during training without the need for a pre-defined negated image-text pair dataset. In each training batch, we identify the most similar image–text pair for every image–text example by computing the cosine similarity between their embedded image features. For each caption, we generate negated variants using a template-based approach, by interacting with another caption in the same batch. Because the negated caption generation relies on the other captions, we can generate diverse and different negated captions in every training epoch. We also propose a negation text-to-image generation benchmark, NEG-TTOI, to evaluate the capability of models to avoid generating undesired objects given negated prompts. In this task, a compositional negated caption is given which contains the desired objects and undesired objects, *e.g., "A women not holding a dog in the car"*. The generative model needs to explicitly recognize what needs to be generated and what should be avoided. We show that our proposed data generation and training pipeline can directly benefit the downstream task of text-to-image generation. Our contributions include:

- We propose a novel and efficient training-time negation generation pipeline, *TNG-CLIP*, to improve CLIP's negation understanding by generating dynamic and diverse negation samples during training without the need for LLMs and pre-defined negation datasets.
- We propose the first benchmark for negation-aware text-to-image generation task, NEG-TTOI, which contains diverse and abundant samples to evaluate model's negation understanding capability.
- We offer extensive experiments to demonstrate that *TNG-CLIP* achieves SOTA performance on diverse negation-aware downstream tasks including image-to-text matching, text-to-image retrieval, and image generation, indicating its robustness across these tasks.

2 Related Works

While recent foundation models, including LLMs and VLMs, have achieved remarkable success across diverse downstream tasks, their ability to handle negation semantics remains limited. In the scope of large-scale foundation models, the study of negation understanding starts from language-only setting, where large language models, instead of vision-language models, are focused. Truong et al. shows the LLM's insensitivity of negation by evaluating SOTA LLMs (Brown et al., 2020; Ouyang et al., 2022; Chung et al., 2022) on diverse text-only negation benchmarks (Hossain et al., 2020; Geiger et al., 2020; Truong et al., 2022). Zhang et al. mentions that scaling-up the size of LLM fails to tackle negation tasks. Also, Varshney et al. analyze and tackle the issue of negation in LLM hallucinations, which also emphasizes the significance of negation understanding in LLMs.

On the other hand, the negation study in VLMs is mainly focused on CLIP (Radford et al., 2021). For example, Quantmeyer et al. conduct experiments and visualize where and how does CLIP model process negation information in each layers. To make CLIP model understand negation, methods (Park et al., 2025; Singh et al., 2024; Alhamoud et al., 2025a) adopt LLMs to generate negation caption, based on existing image-text pair datasets, to fine-tune the CLIP for negation understanding. However, generating million-scale negation caption with LLM is extremely time- and compute-consuming, and the negation caption is associated with fixed negation object. For example, when a image is paired with the negation caption *"A dog not with a boy"*, the word *"boy"* can be substituted with plenty of potentially-existing objects such as *"cat"*, *"ball"*, *"food"* and so on.

Instead of relying on a fixed and stationary dataset throughout the training process, some methods explore the application of dynamic and non-stationary datasets during training process (Wang et al., 2019; Jiang et al., 2024; Böther et al., 2025; Cheng et al., 2025), which is an effective strategy to improve model robustness, generalization, and training efficiency. Inspired by the idea of dynamic dataset training, we generate similar but different negation captions for the same image in every epoch of training, which enhances the diversity of the dataset. Thus, models can learn negation semantics via the absence of multiple negation objects to improve robustness and generalization.

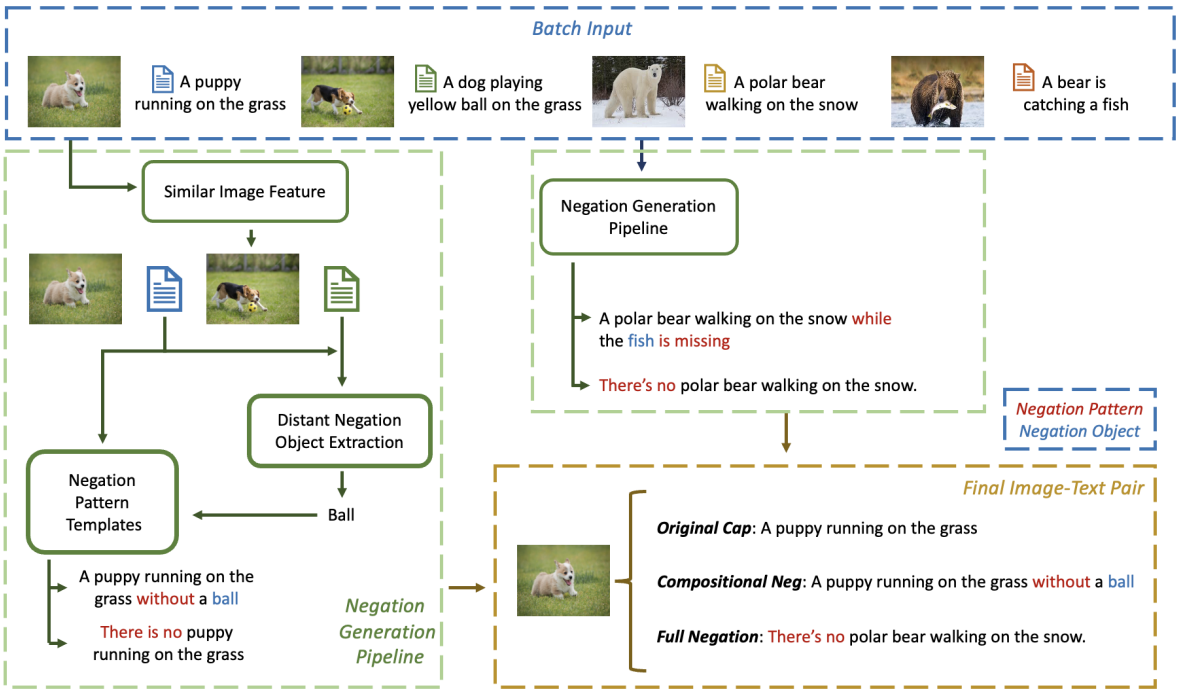

Figure 2: **Training Procedure of *TNG-CLIP***. The diagram shows the data generation pipeline during the training for one sample in the batch. For an image-text pair, P_o, the most similar image pair, P_s is selected by the cosine similarity of their embedded image features. The captions from P_o and P_s are used to find the negation object and generate two types of negation captions. The final image-text set, S_i, for i^{th} image-text pair will be composed of one image, I_i, one original caption, T_{o_i}, one compositional negation caption, T_{nc_i}, and one full negation caption, T_{nf_j} from another random sample.

3 Training-Time Negation Data Generation for Negation Understanding

To make CLIP learn negation semantics with diverse datasets and without the burden of time- and compute-consuming LLM-based negation caption generation, we present our novel training pipeline, **T**raining-Time **N**egation Data **G**eneration for CLIP (*TNG-CLIP*), such that we generate image-text sets with form <I, T_o, T_{nc}, T_{nf}>, from the given image-text pair <I, T_o>, where I and T_o represent the provided image and the original (non-negation) caption in the image-text pair dataset, while T_{nc} and T_{nf} represent the two types of generated negation captions: **compositional negation caption** and **full negation caption**, discussed in Sec 3.1.3.

3.1 Training time data augmentation

We propose a novel negation data-generation pipeline that the negation captions are formed during each batch of training procedure. The negation data generation pipeline for one image in the batch is shown in Figure 2. Overall, for a given image-text pair, P_o, we will first find another similar image-text pair, P_s, select the negation object, O_n, and generate corresponding negation captions, T_{nc} and T_{nf} with the randomly-chosen negation pattern template and form the image-text set, S.

3.1.1 Find similar image-text pairs

To form a semantically reasonable compositional negation caption, T_{nc}, we need to find a proper negation object, O_n, that can be potentially fitted into the original caption, T_o. For example, we want T_{nc} to be "*A dog running with no boy around*", instead of "*A dog running with no whale around*", which is semantically unlikely. Previous methods (Park et al., 2025; Alhamoud et al., 2025b) acquire the proper negation object, O_n, through the reasoning of LLM to find the possible object that might appear in the image but is actually absent. For efficiency, we avoid the use of an LLM, and propose to find the possible O_n of the image-text pair, P_o, from its most similar images-text pair, P_s, in the same batch. Thus, the first step is to find the P_s for every P_o via cosine similarity, between the embedded image features.

Given a visual encoder $E_v(\cdot)$, a batch of images I_b is encoded into the corresponding visual features

$$V_b = E_v(I_b), V_b \in \mathbb{R}^{B \times D}, \quad (1)$$

where B is the batch size and D is the hidden dimension of image feature. For i^{th} image feature, V_{b_i}, we apply cosine similarity

$$V_{bs_i} = \arg\max_{V_j} cos\_sim(V_{b_i}, V_j) \quad (2)$$

to find the most similar image feature, V_{bs_i}, and keep track of the most similar image-text pair, P_{s_i}, associated with the image feature V_{bs_i}.

3.1.2 Select negation object

After having P_s for each image-text pair, P_o, we aim to find the negation object, O_n, exists in P_s's caption that does not exist in the caption of P_o. For caption in P_s, we employ Natural Language Tool Kit (Bird et al., 2009) to extract the POS tag of every word, and only keep those represent nouns. To avoid selecting the object which is too semantically close to the words in original caption and cause conflict, we use WordNet (Miller, 1995) and its hand-curated symbolic network to select the negation object, O_n, with furthest semantics to those words in the original caption.

3.1.3 Template-based negation caption generation

For every T_o and O_n, we employ randomly-chosen negation templates to generate two different types of negation captions: **compositional negation caption**, T_{nc}, and **full negation caption**, T_{nf}. While the compositional negation caption helps model align image with partial negation of a relevant caption, full negation caption makes the image align with the negation semantics of an unrelated caption.

1. **Compositional Negation Caption:** The negation caption is in the format of "*A <negation> B*" , where *A* denotes the original caption, T_o, *B* denotes the negation object, O_n, and *<negation>* represents the negation template that combines the two. For example, let *A* denotes "*A dog playing a ball.*", *B* denotes "*Boy*", and *<negation>* denotes "*There is {caption}, but not a {obj} around.*" The final compositional negation caption, T_{nc}, is "*There is a dog playing a ball, but not a boy around.*" To make the generated captions diverse, we use GPT-4o (OpenAI et al., 2024) to generate 46 different negation patterns.

2. **Full Negation Caption:** The negation caption is in the format of *<negation> A*, which is the negation of the entire caption. We use GPT-4o to generate 18 different negation pattern.

All the negation patterns and the prompt for GPT-4o to generate them are attached in Appendix A.5.

3.1.4 Form new image-text set

Given the original caption, T_o, compositional negation caption, T_{nc}, and full negation caption, T_{nf}, we can now construct the final image-text set, S, for training. For each image I_i, we associate it with the original caption, T_{o_i}, the compositional negation caption, T_{nc_i}, and the full negation caption, T_{nf_j}, $j \neq i$. Please note that we randomly pick the full negation caption, T_{nf_j}, from other image-text pairs, P_j. This is because we want to align the negation of the irrelevant captions to the image and contrast the negation of the relevant caption. Finally, the image-text set, S, is denoated as

$$\text{Image}_i \leftrightarrow \begin{cases} \text{Original}_i \\ \text{Compositional Negation}_i \\ \text{full negation}_j, j \neq i \end{cases}$$

3.2 Asymmetric noise-augmented objective

After negation image-text set generation, each image is associated with three captions, which makes the image-text pair imbalanced. Thus, the image-to-text loss, $\mathcal{L}_{i2t}$, and text-to-image loss, $\mathcal{L}_{t2i}$, become asymmetric. We redefine the functionality of both unidirectional loss to serve different purpose.

Text-to-Image Objective Given that we have three captions for one image, the similarity matrix will be in shape of $3N \times N$, where N denotes the number of the images. We calculate the $\mathcal{L}_{t2i}$ in a single objective by applying same image alignment to the three captions. The text-to-image objective function is defined as:

$$\mathcal{L}_{\text{t2i}} = -\frac{1}{3N} \sum_{j=0}^{3N-1} \log \left(\frac{\exp\left(S_{j,\lfloor \frac{j}{3} \rfloor}/\tau\right)}{\sum_{i=0}^{N-1} \exp\left(S_{j,i}/\tau\right)} \right),$$

where $S_{j,i}$ denotes the similarity between caption j and image i.

Image-to-Text Objective Aligning each image with a negation caption, specifically negation object, is out-of-distribution for pre-trained CLIP because CLIP, which has seen only image–text pairs in which almost all textual components are visually grounded, with no explicit representation of negation. As a result, the pre-trained model struggles to align negation semantics or irrelevant objects with the image. Fine-tuning pre-trained model on such OOD task might lead to worse performance,

because fine-tuning can achieve worse accuracy, by overfitting, when the pretrained models are good and the downstream task distribution shift is large, supported by theory from (Kumar et al., 2022). To solve the above obstacle of overfitting, we introduce label noise to improve the generalization and robustness of the model, inspired by the related works (Rolnick et al., 2018; Xie et al., 2020; Chen et al., 2025). We modified the image-to-text loss such that the text labels are randomly aligned with the image to introduce noise to the objective function. The $\mathcal{L}_{i2t}$ is:

$$\mathcal{L}_{\text{i2t}} = -\frac{1}{N}\sum_{i=0}^{N-1}\log\left(\frac{\exp\left(S_{i,y_i}/\tau\right)}{\sum_{j=0}^{3N-1}\exp\left(S_{i,j}/\tau\right)}\right),$$

where $y_i \sim \mathcal{U}(\{0, 1, \ldots, 3N-1\})$ is a random selected label across all the captions labels.

Combined Objective By introducing noise to $\mathcal{L}_{i2t}$, we only have uni-directional $\mathcal{L}_{t2i}$ helping align negation captions to image. This approach is possible because we freeze the visual encoder during the training, following previous works (Singh et al., 2024; Park et al., 2025). Because the visual encoder is fixed, the visual feature is not updated during image-to-text alignment training, and the model only learn to update text features closer to the pre-trained visual features. The final objective function is then defined as:

$$\mathcal{L} = \frac{1}{2}(\mathcal{L}_{i2t} + \mathcal{L}_{t2i}).$$

The further analysis of the objective function is presented in Appendix A.1.

4 Negation Text-to-Image Generation Benchmark

While negation is an essential part of natural language understanding, a well-designed image generative model should be capable of understanding what to generate and what to avoid. To analyze the generative models' performance on negation prompts, Park et al. proposed negation-aware image generation experiments with only 107 negation prompts, containing simple naive negation pattern of *"no", "not", "without"*. To enable systematic analysis, we design the first negation-based text-to-image generation benchmark, NEG-TTOI, with examples in Table 11. It contains 2000 evaluation samples in the form of <p,q_p,q_n,a_p,a_n>, where p is the prompt mentioning both desired and undesired objects, q_p is positive question about the existence of desired objects, q_n is the negative question about the absence of undesired objects, and a_p and a_n are the answer to q_p and q_n.

4.1 Negation prompts generation pipeline

We follow the procedure of previous works (Park et al., 2025; Alhamoud et al., 2025a) to generate prompts and questions via LLM. We use LLM instead of our negation generation pipeline in Sec 3 because (1) the scale of our evaluation benchmark is much smaller than the scale of training dataset, and (2) we only generate the benchmark prompts and questions once, without the necessity of iterative negation data generation over epochs, which makes the LLM time- and compute-affordable.

We use the MS-COCO Caption (Chen et al., 2015) as the base dataset. The goal of our caption generation pipeline is to transform each caption, which describes the existing scene or objects in the image, into a negation-style caption in which certain elements are explicitly described as absent. To efficiently manipulate the caption with complicated semantics, we leverage GPT-4o (OpenAI et al., 2024) in a multi-step manner from negation prompt generation, evaluation questions generation and quality verification.

1. **Negation Prompt Generation:** For every input caption, we ask LLM to identify a random scene or object that is mentioned in the original caption. The selected scene or object will be used as the negation object to generate negation caption. Once we have the original caption and the negation object, we prompt LLM to rewrite the original caption such that the object should be semantically absent from the original caption.

2. **Evaluation Question Generation** For every negation prompt, we prompt LLM to identify the positive semantics and negative semantics in the sentence while discard the negation pattern. For example, given a negation caption *"A dog playing a yellow ball while there is no man walking around"*, the positive semantics will be *"A dog playing a yellow ball"*, while the negative semantics will be *"man walking around"*. Both the positive semantics and negative semantics are combined with "Is there...?" to form the questions q_p and q_n.

3. **Question Quality Verification** Although GPT-4o is one of the SOTA LLMs for semantic under-

Model	Avg.	Affirmation	Negation	Hybrid	R@5	Neg-R@5
CLIP (Pretrained)	16.28	21.89	16.89	9.99	54.76	47.92
CoN-CLIP	15.70	0.05	36.73	11.97	51.91	48.22
NegCLIP	10.21	9.97	19.76	1.83	**68.73**	**64.41**
CLIP (CC12MNegFull)	46.9	56.49	41.71	42.29	54.20	51.90
TNG-CLIP (Ours)	**52.5**	**68.75**	**44.75**	**43.29**	62.00	61.11

Table 1: Result on Negbench MSCOCO image dataset on image-to-text matching and text-to-image retrieval tasks. **R@5** refers to the Top-5 accuracy on original (non-negation) MSCOCO-Caption dataset, while **Neg-R@5** refers to the Top-5 accuracy on negation MSCOCO-Caption dataset from NegBench.

Model	Avg.	Affirmation	Negation	Hybrid
CLIP (Pretrained)	14.47	31.96	8.34	14.97
CoN-CLIP	22.36	0.01	27.67	24.14
NegCLIP	8.50	22.58	8.62	4.08
CLIP (CC12MNegFull)	52.65	73.75	35.69	62.34
TNG-CLIP (Ours)	**59.23**	**85.92**	**36.39**	**72.80**

Table 2: Result of Negbench image-to-text matching on VOC2007 image dataset

standing, it still might generate text that are semantically incorrect. Thus, verification is necessary to prevent the improper generation. Given the negation prompt, p, positive question, q_p, and negative question q_n, we prompt the LLM to ask whether the semantics in the q_p is stated positively in p, and whether the semantics in the q_n is stated negatively in p with the negation semantics. If the LLM's answer for both questions are correct, the negation data sample will be kept, otherwise it will be discarded.

In the end, NEG-TTOI contains 2000 valid samples, selected from 2500 candidates.

4.2 Evaluation metrics

Unlike image-text matching or retrieval tasks such that the explicit ground truth can be found, evaluating image generation task is relatively subjective. Inspired by (Park et al., 2025; Hu et al., 2023), we employ GPT-4o (OpenAI et al., 2024) to evaluate the existence and absence of the objects. Given a image generated using negation caption as prompt and the positive question and negative question, we evaluate the model's generation quality via the metric of **Compositional Accuracy**: it's **True** if the LLM answers "yes" on positive question and "no" on negative question at the same time.

5 Experiments

To show the capability of our proposed method on multiple downstream tasks, we evaluate our model on negation tasks including image-to-text matching, text-to-image retrieval and text-to-image generation. Our goal is to assess *TNG-CLIP*'s negation semantics understanding via multiple benchmarks and show its generalization and capacity on diverse negation-based scenarios. In the paper, all experiments are performed on a single Nvidia A40 GPU with batch size of 128 and learning rate of 5e-6.

5.1 Matching & retrieval evaluation

To evaluate the negation understanding ability of *TNG-CLIP*, we present the experiments on image-to-text matching and text-to-image retrieval tasks.

Benchmarks We employ the following benchmarks to evaluate the model's performance:

- Valse-Existence (Parcalabescu et al., 2022) benchmark evaluates the model's performance on negation imaget-to-text matching task. Given a image and two text description about the presence and absence of an object in the image, *e.g. "There is animal in the image"/"There is no animal in the image"*, the model should select the best-matched text.

- NegBench (Alhamoud et al., 2025b) benchmark is a comprehensive benchmark to evaluate the negation understanding of models on variant image-to-text matching and text-to-image retrieval tasks. It includes negation-based matching tasks based on both MS-COCO(Chen et al., 2015) and

VOC2007(Everingham et al.) datasets, a text-to-image retrieval task based on MS-COCO evaluation dataset, where the captions are converted into compositional negation style. In the matching task, images are paired with four different captions of three categories: *Affirmation* for *"It include A and B."*, *Negation* for *"Does not include A and B."*, and *Hybrid* for *"Include A but not B."*.

Model	Accuracy
CLIP (Pretrained)	65.16
NegCLIP	73.22
CoN-CLIP	74.15
CLIP (CC12MNegFull)	76.21
NegationCLIP	80.15
TNG-CLIP (Ours)	**81.64**

Table 3: Valse-Existence Image-to-Text Matching

Baselines To evaluate the performance of our method, we compare it against several existing baseline methods for CLIP's negation understanding, including *pretrained-CLIP* (Radford et al., 2021), *NegCLIP* (Yuksekgonul et al., 2023), *CoN-CLIP* (Singh et al., 2024), and CLIP fine-tuned on *CC12M-NegFull* (Alhamoud et al., 2025a). For fair comparison, all of the methods are initialized based on pre-trained CLIP ViT-B/32 model.

Comparison Experiments We present the matching and retrieval task of NegBench-MSCOCO in table 1 and the matching task of NegBench-VOC2007 in table 2. From the tables, we observe that previous methods are lack of generalization on negation-based tasks, but only focus on the negation understanding of specific tasks. For example, *CoN-CLIP*'s performance on matching (affirmation) task is 0.05 and 0.29 on MSCOCO and VOC2007 datasets, which indicates that the method is biased such that it sacrifices the CLIP's performance on non-negation performance for negation improvement. For *NegCLIP*, even though it get the best score on retrieval task, we observe that the affirmation performance is lower than that of the *pretrained-CLIP*, and its performance on matching (hybrid) is low. On the other hand, the *CC12M-NegFull* fine-tuned CLIP presents general improvement of different tasks, indicating its capability of diverse negation tasks. Our method, *TNG-CLIP*, even though slightly underperforms the *NegCLIP* model on retrieval tasks, achieves SOTA performance on all the matching tasks, shows its generalization and high-performance on diverse scenarios.

Strategy	Avg. Acc.
dynamic dataset	51.61 $\pm$ 0.96
fixed dataset	49.52 $\pm$ 1.27

Table 4: Effect of using dynamic dataset. Evaluation on NegBench-MSCOCO image-to-text matching task.

Similarly, the evaluation on Valse-Existence dataset, in Table 3, further proofs *TNG-CLIP*'s, capability of negation understanding. While the benchmark is first used by *NegationCLIP* (Park et al., 2025) and achieves promising result of 80.15 on CLIP ViT-B/32 based models, our method gets better performance, 81.64, which is higher than all other negation-understanding CLIP baselines.

Effectiveness of Dynamic Dataset The training-time data generation pipeline generates the negation caption based on the other image-text pairs in the same batch, which makes the negation caption of same image different in every epoch. We analyze the effect of such dynamic dataset and compare how the performance differs from using fixed dataset. We store the image-text set, S, generated in each training epoch for every epoch as the fixed dataset. We then use the fixed dataset to replace the data generation pipeline to fine-tune the CLIP model. To get statistically significant comparison result, we repeat the *TNG-CLIP*'s training procedure for 10 times and use 10 fixed dataset collected from different training epochs to fine-tune pre-trained CLIP with same objective function and hyper-parameters. We present the mean and standard deviation in Table 4. We observe that the performance of *TNG-CLIP* is higher than using fixed dataset, and the standard deviation is also smaller than the fixed one. We explain such phenomenon as the CLIP's fine-tuning on fixed dataset constrains the model's negation understanding to specific <*caption, negation object*> pair, thus harms the generalization of the model on negation tasks, leading to lower mean accuracy. At the same time, the data variance among every epoch for *TNG-CLIP* works as a natural regularization to prevent overfitting and memorizing incorrect correlation, thus lead to smaller standard variance.

More analytic experiments are in Appendix A.4 and A.3.

5.2 Text-to-Image Generation

Model	Arch.	Acc.
SD-1.5	ViT-L/14	32.60
SDXL-1.0	ViT-L/14	27.45
SD-1.5 w/ CoN-CLIP	ViT-L/14	28.40
SD-1.5 w/ TNG-CLIP (ours)	ViT-L/14	**45.65**
pretrained-CLIP + proj	ViT-B/32	28.25
NegCLIP + proj	ViT-B/32	33.85
CoN-CLIP + proj	ViT-B/32	24.05
CC12MNegFull + proj	ViT-B/32	36.95
TNG-CLIP + proj (ours)	ViT-B/32	**41.70**

Table 5: Image Generation on NEG-TTOI benchmark

5.2.1 CLIP for Image Generation Task

Although CLIP model is mostly used to do image-text matching tasks, it can be applied to text-to-image generation tasks indirectly. For example, the text encoder from stable diffusion model is the original copy of CLIP ViT-L/14's text encoder (Rombach et al., 2022). To evaluate the negation understanding of CLIP in text-to-image generation field, Park et al. provides a simple yet effective way, by replacing the original text encoder from stable diffusion model with their proposed negation-aware CLIP. This direct substitution is possible because they fine-tune only the text encoder, preserving the original image embedding space and maintaining the text feature alignment with it.

5.2.2 Experiment Setup

Following the strategy mentioned above, we fine-tuned our *TNG-CLIP* from pretrained CLIP ViT-L/14 model, and replace the original stable diffusion model's text encoder with ours.

However, most baseline methods are fine-tuned only on CLIP ViT-B/32 model, it is difficult to do the direct substitution due to the mismatch of output feature dimension. To tackle such issue, we attach a MLP projector after the frozen text encoder, and perform knowledge distillation between CLIP ViT-L/14's text encoder acts and CLIP ViT-B/32's text encoder with projector, to align the output of projected CLIP ViT-B/32 text encoder similar to that of CLIP ViT-L/14 text encoder. We perform add MLP to all the baseline methods and fine-tune the MLP, with text encoder frozen, on the same dataset, MS-COCO Caption (Chen et al., 2015).

5.2.3 Experiment Analysis

The comparison results on NEG-TTOI benchmark are presented in Table 5. The upper table shows the comparison with CLIP ViT-L/14's text encoder architecture. We choose SD-1.5 (Rombach et al., 2022) as the generative model backbone and replace its text encoder with that of ours and *CoN-CLIP*'s. All the experiment here are the zero-shot performance on NEG-TTOI benchmark. We observe that among the all, using our *TNG-CLIP*'s text encoder achieves the best accuracy, indicating its outstanding capability of handling negation feature for image generation. On the other hand, the accuracy of *CoN-CLIP* is lower than original stable diffusion model, which shows its deficiency on image generation task.

The lower table presents the accuracy of SD-1.5 by replacing its text encoder with the combination of CLIP ViT-B/32 based architecture and the fine-tuned MLP projector. Noticing that the accuracy of our method using CLIP ViT-B/32's text encoder is 41.70, while that for using CLIP ViT-L/14's text encoder is 45.65, showing that the projected ViT-B/32 text encoder is not as effective as ViT-L/14's text encoder, and is only used for the purpose of providing accessible and fair comparison between the baselines on image generation task. Among the all, our method's text encoder still achieves the best accuracy, and the clip fine-tuned with *CC12MNegFull* (Alhamoud et al., 2025a) is the second best, similar with its performance in image-text matching tasks.

We provide more detailed image generation task analysis in Appendix A.2.

6 Discussion & Conclusion

In this paper, we focus on the critical problem of improving negation understanding for CLIP. Instead of using pre-generated fixed negation dataset, we propose a training-time negation data generation pipeline to generate dynamic negation caption during the training time, addressing the time- and compute- inefficiency problem of previous dataset. We also show that using dynamic negation caption during the training can improve mdoel's generalization and boost the performance of negation fine-tuned CLIP. On the other hand, we propose the first negation-aware text-to-image generation evaluation benchmark to expand the horizon of negation-related benchmarks. Overall, our work underscores the negation understanding in the study of vision language model, and call for the wider exploration of negation-aware model in diverse tasks.

7 Limitations

In this paper, we propose a negation-aware CLIP, *TNG-CLIP*, trained via the novel efficient training-time negation data generation pipeline. We also propose a negation text-to-image generation benchmark, NEG-TTOI, to evaluate the capability of generative model's performance with negation semantics. However, although we have shown the performance and generalization of *TNG-CLIP* via multiple benchmarks, we see the limit of our paper:

- In the paper, we mainly focus on the negation understanding of CLIP model. As the lack of negation understanding is an overall challenge among all vision language models, further exploration on negation-awareness of diverse VLMs is necessary.
- The training-time negation data generation pipeline is currently limited to image-text pair dataset, which is adopted to apply contrastive learning. Our negation data generation pipeline has the potential to be extended beyond image-text pairs, eg. visual question answering dataset, thus supports the negation-awareness training with objective function other than contrastive loss.

References

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip Torr, Yoon Kim, and Marzyeh Ghassemi. 2025a. Vision-language models do not understand negation. *Preprint*, arXiv:2501.09425.

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip Torr, Yoon Kim, and Marzyeh Ghassemi. 2025b. Vision-language models do not understand negation. *Preprint*, arXiv:2501.09425.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Maximilian Böther, Ties Robroek, Viktor Gsteiger, Robin Holzinger, Xianzhe Ma, Pınar Tözün, and Ana Klimovic. 2025. Modyn: Data-centric machine learning pipeline orchestration. *Proceedings of the ACM on Management of Data*, 3(1):1–30.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *Preprint*, arXiv:2102.08981.

Hao Chen, Zihan Wang, Ran Tao, Hongxin Wei, Xing Xie, Masashi Sugiyama, Bhiksha Raj, and Jindong Wang. 2025. Impact of noisy supervision in foundation model learning. *Preprint*, arXiv:2403.06869.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *Preprint*, arXiv:1504.00325.

Ziheng Cheng, Zhong Li, and Jiang Bian. 2025. Data-efficient training by evolved sampling.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Dumitru, Ian Goodfellow, Will Cukierski, and Yoshua Bengio. 2013. Challenges in representation learning: Facial expression recognition challenge. `https://kaggle.com/competitions/challenges-in-representation-learning-facial-expression-` Kaggle.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 163–173, Online. Association for Computational Linguistics.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: data, models and evaluation metrics. *J. Artif. Int. Res.*, 47(1):853–899.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. 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 (EMNLP)*,

pages 9106–9118, Online. Association for Computational Linguistics.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *Preprint*, arXiv:2303.11897.

Yiding Jiang, Allan Zhou, Zhili Feng, Sadhika Malladi, and J Zico Kolter. 2024. Adaptive data optimization: Dynamic sample selection with scaling laws. *arXiv preprint arXiv:2410.11820*.

Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. *Preprint*, arXiv:2202.10054.

George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 8253–8280. Association for Computational Linguistics.

Junsung Park, Jungbeom Lee, Jongyoon Song, Sangwon Yu, Dahuin Jung, and Sungroh Yoon. 2025. Know "no" better: A data-driven approach for enhancing negation awareness in clip. *Preprint*, arXiv:2501.10913.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2016. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Preprint*, arXiv:1505.04870.

Vincent Quantmeyer, Pablo Mosteiro, and Albert Gatt. 2024. How and where does clip process negation? *Preprint*, arXiv:2407.10488.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. 2018. Deep learning is robust to massive label noise. *Preprint*, arXiv:1705.10694.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. *Preprint*, arXiv:2112.10752.

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

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not naysayers: An analysis of language models on negation benchmarks. *Preprint*, arXiv:2306.08189.

Thinh Hung Truong, Yulia Otmakhova, Timothy Baldwin, Trevor Cohn, Jey Han Lau, and Karin Verspoor. 2022. Not another negation benchmark: The NaN-NLI test suite for sub-clausal negation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 883–894, Online only. Association for Computational Linguistics.

Neeraj Varshney, Satyam Raj, Venkatesh Mishra, Agneet Chatterjee, Ritika Sarkar, Amir Saeidi, and Chitta Baral. 2024. Investigating and addressing hallucinations of llms in tasks involving negation. *Preprint*, arXiv:2406.05494.

Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2019. Dynamic sentence sampling for efficient training of neural machine translation. *Preprint*, arXiv:1805.00178.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. 2020. Self-training with noisy student improves imagenet classification. *Preprint*, arXiv:1911.04252.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? *Preprint*, arXiv:2210.01936.

Yuhui Zhang, Michihiro Yasunaga, Zhengping Zhou, Jeff Z. HaoChen, James Zou, Percy Liang, and Serena Yeung. 2023. Beyond positive scaling: How negation impacts scaling trends of language models. *Preprint*, arXiv:2305.17311.

A Appendix

A.1 Ablation Study of Asymmetric Noise-Augmented Objective

In order to train the negation-aware CLIP for diverse tasks, we propose a novel asymmetric noise-augmented loss that different from the original contrastive loss of CLIP. We exam the contribution of each component in this novel objective function with analytic ablation study. Within the objective function, we split its component to four parts based on the functionality of each:

- **compositional alignment** refers to align the compositional negation caption to the image in $\mathcal{L}_{t2i}$.
- **full alignment** refers to align the full negation caption to the image in the $\mathcal{L}_{t2i}$.
- **original alignment** refers to align the original caption to the image in $\mathcal{L}_{t2i}$.
- **noise alignment** refers to align the random-chose caption to the image in $\mathcal{L}_{i2t}$.

The analytic results are presented in Table 6, with the evaluation on Negbench-MSCOCO matching and Negbench-MSCOCO Retrieval tasks. From the table, we observe that by eliminating compositional negation, full negation and original caption from $\mathcal{L}_{t2i}$ separately, the corresponding performance in matching task drops. For example, without original caption, the affirmation accuracy drops from 68.75 to 60.18. At the same time, the accuracy of negation retrieval tasks remains similar, indicating that the components in $\mathcal{L}_{t2i}$ are not the primary factors for it.

We then analyze the effect of random noise in $\mathcal{L}_{i2t}$. Instead of letting image random choose caption, we match the image to its corresponding original, compositional negation and full negation captions as three independent experiments. Additionally, we let images to randomly match one of its corresponding original, compositional negation and full negation caption, and even directly delete the $\mathcal{L}_{i2t}$ loss. Through the experiments, we found that **without the random noise, performance of retrieval task drops significantly.** This matches the hypothesis we proposed in Sec 3.2 that negation dataset is an OOD task for pre-trained CLIP, the direct fine-tuning may cause worse performance.

Lastly, we propose and examine another loss objective: can we split the generated image-text set, S, to form three image-text pairs for each of original, compositional negation and full negation, and apply normal contrastive loss on the three independently? We implement such objective function and present it at the bottom of the table. We observe that by doing so, the performance of both matching task and retrieval task are sub-optimal. While there is also no noise label added to the objective training, the worse result on using independent losses, again, emphasizes the importance of adding noise when fine-tuning CLIP on negation-related dataset.

A.2 More Analysis of Image Generation Experiment

In the image generation task, we observe the inefficiency of original Stable Diffusion and CoN-CLIP in the NEGTTOI benchmark. But why this happens? To further explore that, we evaluate the performance of models with two analytic metrics: **Positive Accuracy** and **Negative Accuracy**. Given a prompt "generate A without B", **Positive Accuracy** measures if the image contains A, and **Negative Accuracy** measures if the image doesn't contain B. The result is presented in Table 7. In the table, we can observe that for original Stable Diffusion model, the positive accuracy is higher than that of using our method or CoN-CLIP as text encoder, but the negative accuracy is much lower. This explicitly shows that the original text encoder cannot process negation semantics to help avoid the generation of unwanted objects. On the other hand, adopting CoN-CLIP as text encoder can significantly boost the negative accuracy, but at the same time, its performance on positive accuracy becomes low. This indicates the CoN-CLIP model is a biased model towards negation-understanding, while ignores the generalization on other non-negation tasks.

A.3 Non-Negation Generalization on Image Classification

Although TNG-CLIP is specifically designed for negation understanding, it is important to ensure that its performance on non-negation tasks remains intact, in another word, it should not suffer from catastrophic forgetting on tasks that the original pre-trained CLIP model was capable of handling. Inspired by the experiments from (Singh

Model	Avg.	Affirmation	Negation	Hybrid	Neg-R@5
TNG-CLIP	52.50	68.75	44.75	43.29	61.11
Ablation of Caption Category					
w/o compositional	48.15	65.45	38.02	40.10	56.66
w/o full	51.31	75.63	24.75	51.44	60.79
w/o original	46.66	60.18	45.82	37.79	59.91
Ablation of Noise					
$\mathcal{L}_{i2t}$: original	52.49	81.93	16.09	56.66	45.32
$\mathcal{L}_{i2t}$: compositional	50.13	78.05	11.97	57.40	45.39
$\mathcal{L}_{i2t}$: full	40.29	45.12	44.11	31.85	48.58
$\mathcal{L}_{i2t}$: random of three	47.92	58.66	43.26	41.10	50.66
w/o $\mathcal{L}_{i2t}$	46.24	59.54	36.04	42.29	50.49
independent losses	46.49	55.41	43.74	40.05	50.19

Table 6: Ablation Study on NegBench MSCOCO matching task

Model	Arch.	Positive	Negative
SD-1.5	ViT-L/14	80.85	41.95
SDXL-1.0	ViT-L/14	87.05	32.30
SD-1.5 w/ CoN-CLIP	ViT-L/14	46.25	72.50
SD-1.5 w/ TNG-CLIP (ours)	ViT-L/14	75.80	63.05
pretrained CLIP + proj	ViT-B/32	45.65	67.25
NegCLIP + proj	ViT-B/32	67.80	52.10
CoN-CLIP + proj	ViT-B/32	39.45	72.65
CC12MNegFull + proj	ViT-B/32	53.76	71.80
TNG-CLIP + proj	ViT-B/32	63.65	68.50

Table 7: Image Generation on Neg-TtoI benchmark

et al., 2024), we conduct the zero shot image classification on TNG-CLIP and pre-trained CLIP with eight diverse benchmarks: **FER2013** (Dumitru et al., 2013), **Flickr-8K** (Hodosh et al., 2013), **Flickr-30K** (Plummer et al., 2016), **MS-COCO** (Chen et al., 2015), **SUN397** (Xiao et al., 2010), **VOC2007** (Everingham et al.), **CIFAR-10** (Krizhevsky, 2009), **CIFAR-100** (Krizhevsky, 2009). The top1 and top5 accuracy score is presented in Figure 3. In the figure, we observe that the zero-shot performance of TNG-CLIP remains similar with that of pre-trained CLIP, indicating there is no catastrophic forgetting or overfitting to our proposed method. Surprisingly, we also observe that in some cases, such as Flickr-8K, Flickr-30K, MS-COCO and VOC2007 benchmarks, the TNG-CLIP outperforms the pre-trained CLIP, illustrating improving the negation understanding can improve model's performance on general tasks.

A.4 Time-Efficiency Test

As we generate data samples during the training stage, does the generation pipeline significantly slower the training process and becomes time-consuming? We compares the average training time per batch on the same GPU device, Nvidia-A40, with and without the data generation pipeline in Table 8. For every batch, the data generation pipeline takes 0.13 sec, which is only 2.55% slower than without using the data generation pipeline. Thus, adding the data generation pipeline to the training is still time-efficient.

A.5 Template-based Negation Pattern

During the negation caption generation, we use pre-defined LLM-generated negation pattern template to convert the original caption and negation object

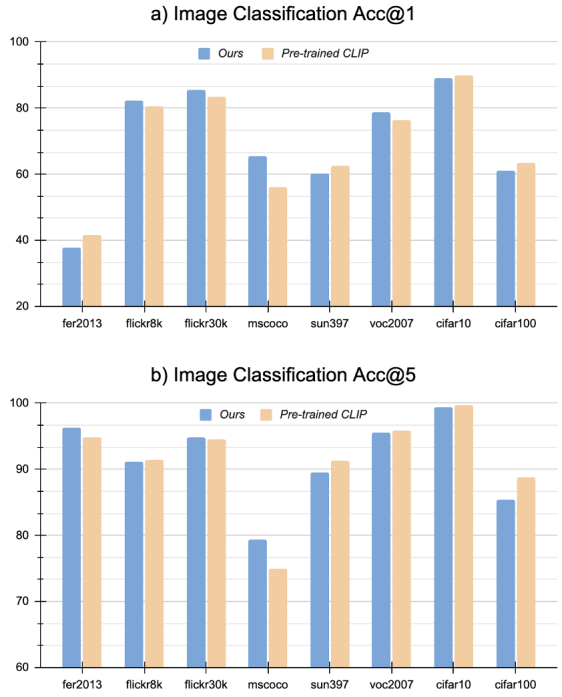

Figure 3: The zero shot image classification accuracy of pre-trained CLIP and TNG-CLIP on eight image classification benchmarks.

Strategy	Time (sec)
w/o data generation	4.97
w/ data generation	5.10
data generation	0.13

Table 8: Time Efficiency for Data Generaiton

to compositional negation caption and full negation caption. We present the template we used here in Table 9 and Table 10.

There's no {cap} in the image.	No {cap} is included in the image.
There is not {cap} in the image.	The image does not have {cap}.
No {cap} is present in the image.	{cap} is not present in the image.
{cap} is absent.	No {cap} is present.
There isn't any {cap}.	Not a single {cap} can be seen.
The image is without {cap}.	The image is lacking {cap}.
There appears to be no {cap} in the image.	The image does not contain {cap}.
There does not exist {cap} in the image.	There is nothing about {cap}.
There isn't any {cap}.	No {cap} is seen in the image.

Table 9: Templates for full negation caption generation, we replace the cap with the provided original caption.

{cap} with no {obj}.	{cap} without {obj}
{cap} that do not have {obj}.	{cap} having no {obj}.
{cap} not include {obj}.	{cap} excluding {obj}.
{cap}, but no {obj} are present.	{cap}, though no {obj} can be seen.
{cap} without any {obj} in sight.	{cap} yet no {obj} are nearby.
{cap} but no {obj} are visible.	{cap} and no {obj} are anywhere around.
{cap}, without any {obj} in the vicinity.	{cap}, with no {obj} in the surroundings.
{cap}, but no {obj} are in the area.	{cap}, and no {obj} can be found nearby.
{cap} in the absence of {obj}.	{cap}, where no {obj} are present.
{cap} with an absence of {obj}.	{cap}, as no {obj} are around.
{cap}, while lacking any {obj}.	{cap} but no {obj} are engaging.
{cap} with no {obj} participating.	{cap} yet no {obj} are interacting.
{cap}, as no {obj} are involved.	{cap}, while {obj} remain absent from the scene.
{cap} though no {obj} can be spotted.	{cap} where no {obj} are noticeable.
{cap} but no {obj} are detectable.	{cap}, as no {obj} are apparent.
{cap}, with no sight of any {obj}.	No {obj} is visible, but {cap}.
No {obj} can be seen, while {cap} happens.	No {obj} is present, yet {cap} continues.
No {obj} appears in sight, but {cap} unfolds.	Not a single {obj} is noticeable, but {cap}.
No trace of {obj} can be found, while {cap} occurs.	No sign of {obj} is apparent, but {cap} is happening.
There is no {obj} in view, but {cap} takes place.	None of the {obj} are around, yet {cap} continues.
Not even one {obj} is nearby, but {cap} is ongoing.	No {obj} exists in the scene, while {cap} happens.
Absolutely no {obj} is here, yet {cap} remains.	Nowhere can {obj} be found, but {cap} is evident.
Nowhere in sight is any {obj}, yet {cap} unfolds.	No {obj} is around in the surroundings, but {cap} is occurring.

Table 10: Templates for compositional negation caption generation, we replace the cap with the provided original caption and obj with the corresponding negation object.

compositional negation caption	positive question	negative question
A room painted in blue with a white sink, but no door.	Is there a room painted in blue with a white sink?	Is there a door?
A shot inside a kitchen without anyone present.	Is there a kitchen shown?	Is there anyone present?
A woman is walking on the sidewalk without her dog.	Is there a woman walking on the sidewalk?	Is there her dog?
A man without a bike at a marina.	Is there a man at a marina?	Is there a bike?
A man is sitting on a bench without a bicycle nearby.	Is there a man sitting on a bench?	Is there a bicycle nearby?
There's no kitchen sink beside the door and countertop.	Is there a door and countertop?	Is there a kitchen sink beside the door and countertop?
A bathroom without a checkered black and white tile floor.	Is there a bathroom?	Is there a checkered black and white tile floor?
A house boat is moored on a riverbank with no bikes in sight.	Is there a house boat moored on a riverbank?	Is there a bike?
A train missing a striped door waiting on a train track.	Is there a train waiting on a train track?	Is there a striped door?
A small airplane flying without a jet nearby.	Is there a small airplane flying?	Is there a jet nearby?
A woman is seen without a horse in front of a fence with razor wire.	Is there a woman in front of a fence with razor wire?	Is there a horse?
No vans are traveling over a bridge next to train tracks.	Is there a bridge next to train tracks?	Is there a van?
A person riding a bicycle without any river nearby.	Is there a person riding a bicycle?	Is there a river nearby?
No giraffes can be seen in the wood and metal fenced enclosure.	Is there a wood and metal fenced enclosure?	Is there a giraffe?
A row team without a lead woman shouting.	Is there a row team?	Is there a lead woman shouting?
A lady is sitting in a room devoid of any bright pink walls.	Is there a lady sitting in a room?	Is there a bright pink wall?
A man carrying a plate without any food on it.	Is there a man carrying a plate?	Is there any food on the plate?

Table 11: Example from *Neg-TtoI* negation image generation benchmark