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 approach 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. We overcome these limitations by (1) introducing a training-time negation data generation pipeline for CLIP fine-tuning such that large-scale negation captions are efficiently generated during training, which only increases 2.8% extra training time, and (2) proposing the first benchmark, NEG-T2I, for evaluating text-to-image generation models on prompts containing negation to assess the 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 the proposed NEG-T2I.

1 Introduction

Vision-language models (VLMs), such as CLIP (Radford et al., 2021), provide an efficient approach for tackling 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 with a ball.", "There is **no** tree on the street."* Negation is a fundamental aspect of human reasoning, enabling precise descriptions of constraints and expectations in communication. Without proper negation understanding, VLMs may generate and retrieve semantically incorrect content, particularly in complicated scenarios where the presence or absence of specific elements critically alters meaning.

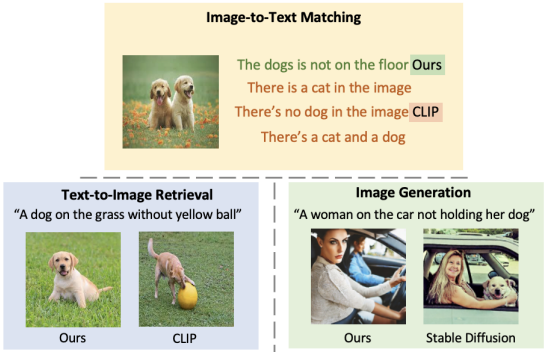

Figure 1: We present *TNG-CLIP*, a negation-aware CLIP model that excels at image-to-text matching, text-to-image retrieval, and a newly proposed image generation benchmark (NEG-T2I).

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, where each image sample is paired with a corresponding negation caption, 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. Given that existing vision-language datasets (Chen et al., 2015; Changpinyo et al., 2021) contain millions of samples, generating the negation dataset is highly time- and compute-intensive. (2) Unlike standard semantic descriptions, which are typically grounded in observable features, the negation process introduces arbitrariness by specifying absent concepts that are not visually present. For example, given an image of *"a dog playing with a ball"*, one could construct multiple valid negation captions such as *"a dog playing with a ball while no man is present"* or *"a dog playing with a ball but not on the beach"*. By generating fixed negation captions, previous methods may constrain the diversity of negation sce-

narios, thus limiting the generalization of the fine-tuned VLMs on negation understanding tasks. (3) Previous methods are mainly evaluated on image-to-text matching and text-to-image retrieval tasks. To maintain the versatility of CLIP-style joint embeddings, however, evaluation should extend beyond matching-based tasks to include generation-based tasks, where the text encoder can serve as part of a generative model (Rombach et al., 2022).

Thus, we propose a new data generation and training pipeline which efficiently 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 sample by computing the cosine similarity between their embedded image and text features. For each caption, we generate several synthetic negated captions 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 show in our experimental results that the synthetic template-based captions, even though may not be as natural and fluent as human-written captions, can significantly improve the model's negation understanding capacity. We also propose a negation text-to-image generation benchmark, NEG-T2I, 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 woman 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. Overall, 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 large-scale benchmark for negation-aware text-to-image generation task, NEG-T2I, which contains diverse and abundant samples to evaluate models' 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.

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 settings, where the focus is on large language models rather than vision-language models. Truong et al. demonstrate that LLMs are insensitive to 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. mention 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, thereby emphasizing the significance of negation understanding in LLMs.

On the other hand, the study of negation in VLMs is mainly focused on CLIP (Radford et al., 2021). For example, Quantmeyer et al. conduct experiments and visualize where and how the CLIP model processes negation information in each layer. To encourage processing negation semantics, methods (Park et al., 2025; Singh et al., 2024; Alhamoud et al., 2025a) adopt LLMs to generate negation captions based on existing image–text pair datasets to fine-tune CLIP for negation understanding. However, generating million-scale negation captions with LLM is highly time- and compute-intensive, and the negation captions are associated with fixed negation objects. For example, when an image is paired with the negation caption *"A dog not with a boy,"* the word *"boy"* can be substituted with an arbitrary noun (*"cat," "ball," "food"*), making covering the space via negation captions alone infeasible.

Instead of relying on a fixed and stationary dataset throughout training, some methods explore the application of dynamic and non-stationary datasets (Wang et al., 2019; Cai et al., 2023; Jiang et al., 2024; Böther et al., 2025; Cheng et al., 2025), which is an effective strategy to improve model ro-

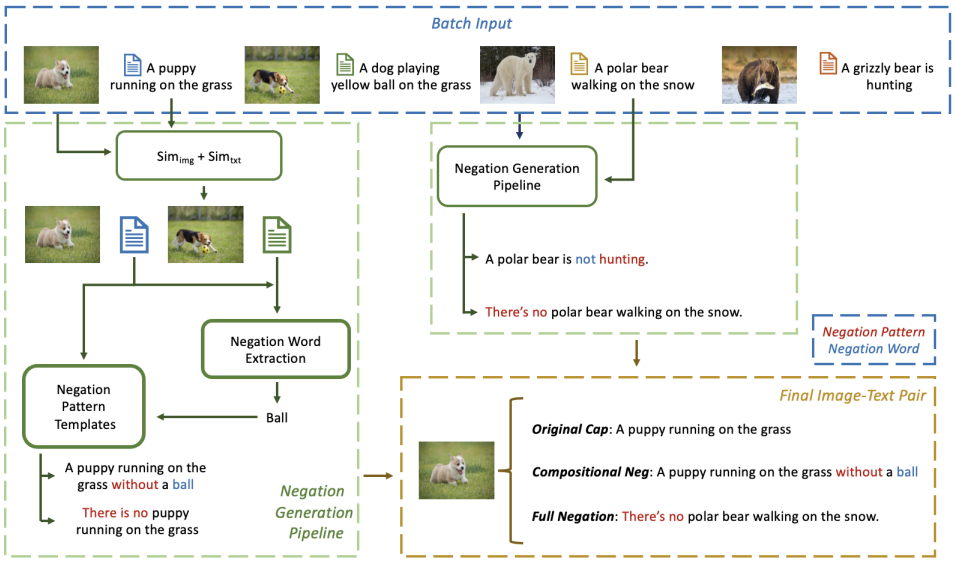

Figure 2: **Training Procedure of *TNG-CLIP***. The diagram shows the data generation pipeline during the training for one sample in the batch. For an image-text pair, P_i, the most similar image pair, $P_{s(i)}$, is selected by the cosine similarity of their embedded image and text features. The captions from P_i and $P_{s(i)}$ are used to find the negation word 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, C_i^o, one compositional negation caption, C_i^c, and one full negation caption, C_j^f from another random sample.

bustness, generalization, and training efficiency. Inspired by the idea of dynamic dataset training, we generate varied negation captions for the same image in each training epoch, thereby enhancing dataset diversity. Thus, models can learn negation via the absence of multiple negation objects to improve performance and generalization.

3 Training-Time Negation Data Generation for Negation Understanding

To enable CLIP to learn negation semantics without relying on expensive LLM-generated datasets, we propose **T**raining-Time **N**egation Data **G**eneration for CLIP (*TNG-CLIP*). Our method dynamically constructs negated captions during training, ensuring semantic diversity while avoiding the inefficiencies of static pre-computed datasets. Given an image–caption pair $P_i = (I_i, C_i^o)$, where C_i^o denotes the original caption, we generate two additional negated captions: a **compositional negation caption** C_i^c and a **full negation caption** C_i^f. The resulting expanded training instance for image I_i is

$$S_i = \{I_i, C_i^o, C_i^c, C_i^f\}.$$

3.1 Training time data augmentation

Our goal is to efficiently generate large-scale negated captions that are semantically diverse. While the generated captions are not intended to match the linguistic fluency of human-annotated captions, they provide a form of large-scale supervision with substantial semantic diversity. Our experimental results show that such captions are highly effective for enhancing the model's ability to understand negation. 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 design a pipeline to find the most similar image-text pair, select the negation word and construct the negation caption with negation templates.

3.1.1 Find similar image-text pairs

To produce a logical negation caption, we need candidate concepts that could reasonably appear in the given image context but are in fact absent. For example, a caption may describe a dog "*not playing with a ball*" or "*not accompanied by a boy*," but is less likely to state that the dog is "*not flying an airplane*." Therefore, for each image-text pair P_i, we first identify a semantically similar image–text pair, $P_{s(i)}$, within the batch to serve as a source of candidate negation words.

Let $f_v(\cdot)$ and $f_t(\cdot)$ denote the image and text encoders. For a batch of images $I = \{I_i\}_{i=1}^B$ and captions $T = \{T_i\}_{i=1}^B$, we obtain

$$E_i^v = f_v(I_i),\ E_i^t = f_t(T_i),\ E_i^v, E_i^t \in \mathbb{R}^D, \quad (1)$$

where B is the batch size and D is the hidden dimension of image and text features. We define the cosine similarity between two instances i and j as

$$s_{ij}^v = cos(E_i^v, E_j^v),\ s_{ij}^t = cos(E_i^t, E_j^t), \quad (2)$$

where $j \in \{0, 1, ..., B\}, j \neq i$. To jointly capture the semantic proximity in both modalities, we compute the combined similarity as

$$s_{ij} = s_{ij}^v + s_{ij}^t. \quad (3)$$

The most similar pair to P_i is then selected as

$$s(i) = \arg\max_{j \neq i} s_{ij}. \quad (4)$$

This nearest neighbor $P_{s(i)}$ provides candidate objects or attributes that can be used to form diverse negation captions for P_i.

3.1.2 Select negation word

Once we identify a semantically related pair, we must determine which word can serve as a negation candidate. We denote that negation word as W_n. For the caption $P_{s(i)}$, we extract candidate words using part-of-speech tagging with Natural Language Tool Kit (Bird et al., 2009), and only focus on nouns, verbs, and adjectives that do not exist in P_i's caption. To maintain semantic consistency of verbs and adjectives, only the words sharing the same head noun as C_i^o are chosen. For example, given the caption from P_i as "*a dog is playing on the floor*," and the caption from $P_{s(i)}$ as "*a dog is sleeping*," the negation verb can be "*sleeping*." However, if the caption from $P_{s(i)}$ is "*a girl is dancing on the floor*," then there will be no negation verb selected. Finally, one candidate is then randomly sampled as W_n.

3.1.3 Template-based negation caption generation

After selecting the W_n, we will convert it into a fluent negated caption. To enhance the diversity of the generated caption, we combine the W_n with a randomly-chosen negation template, from a negation template list, to generate two different types of negation captions: **compositional negation caption**, C^c, and **full negation caption**, C^f.

1. **Compositional Negation Caption for Noun-based Negation Word:** The negation caption is in the format of "*A <negation> B*," where *A* denotes the original caption, C^o, *B* denotes the negation object, W_n, and <*negation*> represents the negation template that combines the two. For example, let *A* denotes "*A dog playing with a ball,*" *B* denotes "*boy,*" and <*negation*> denotes "*There is {caption}, but not a {obj} around.*" The final compositional negation caption, C^c, is "*There is a dog playing with 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 for noun-negation words.

2. **Compositional Negation Caption for Verb- & Adjective-based Negation Word:** To avoid conflicts between the W_n and the existing verbs and adjectives in P_i, we replace the original verb or adjective word with the W_n, instead of directly plugging that word into the sentence. For example, if the original caption is "*there is a red apple,*" and W_n is "*green*," we form the new caption to be "*there is a non-green apple*." Similarly, if the original caption is "*a boy is crying*" and W_n is "*sleeping*," we form the new caption to be "*a boy is not sleeping*."

3. **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 patterns.

The negation patterns are attached in the Appendix A.6.

For each image I_i, we construct $S_i = \{I_i, C_i^o, C_i^c, C_i^f\}$. Please note that for C_i^f, we randomly pick the full negation caption from other image-text pairs. We want to align the negation of the irrelevant captions to the image and contrast the negation of the relevant captions.

3.2 Asymmetric noise-augmented objective

Since each image is paired with three captions, the CLIP's original symmetric image-to-text loss, $\mathcal{L}_{i2t}$, and text-to-image loss, $\mathcal{L}_{t2i}$, become imbalanced and asymmetric. We redefine the CLIP's contrastive loss such that the functionality of both unidirectional losses serve different purposes.

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

three captions. The 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 a negation object, is out-of-distribution for pre-trained CLIP, which was trained on 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 a pre-trained model on such an 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, which is supported by the theory from (Kumar et al., 2022). Inspired by related works (Rolnick et al., 2018; Xie et al., 2020; Chen et al., 2025), to solve the above obstacle of overfitting, we introduce label noises to improve the generalization and robustness of the model. We modify $\mathcal{L}_{i2t}$ that the text labels are aligned with a random image to introduce noises 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 noises to $\mathcal{L}_{i2t}$, we only have uni-directional $\mathcal{L}_{t2i}$ helping align negation captions to image. This strategy is viable because we freeze the image encoder during training, following previous works (Singh et al., 2024; Park et al., 2025). Because the image encoder is fixed, its output image feature for the same image stays fixed during image-to-text alignment training, and the model only learns to update the text feature closer to the pre-trained image features. The final objective function is then defined as:

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

4 Negation Text-to-Image Generation Benchmark

Negation is essential to natural language understanding, and a generative image model should be capable of understanding what to avoid in the presence of negation captions. 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 and short phrases, *(e.g., "A man not wearing a hat")*, with limited negation patterns such as *"no," "not," "without."* To enable systematic analysis, we design the first large-scale negation-based text-to-image generation benchmark, NEG-T2I, with examples in the Appendix A.6. It contains 2000 evaluation samples in the form of <p, q_p, q_n, a_p, a_n>, where p is the full-sentence caption mentioning both desired and undesired objects, q_p is the positive question asking the existence of desired objects, q_n is the negative question asking the absence of undesired objects, and a_p and a_n are the answers to the q_p and the 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 the LLM. We use the LLM instead of our negation generation pipeline in Sec 3 because (1) the scale of our evaluation benchmark, 2000, is much smaller than the scale of training dataset, hundred-thousands, which makes the LLM-based approach computationally affordable, and (2) we need the reasoning capacity from the LLM to accurately identify the desired objects and the undesired objects to form the q_p and the q_n, which exceeds the capacity of our generation pipeline in Sec 3.

We derive the NEG-T2I from the image-text pairs of the MS-COCO Caption (Chen et al., 2015). The goal of our caption generation pipeline is to transform each caption into a negation-style caption. To efficiently manipulate the caption with complicated semantics, we leverage GPT-4o (OpenAI et al., 2024) in a multi-step approach to construct the negation caption, the steps include negation prompt generation, evaluation questions generation, and quality verification.

4.1.1 Negation prompt generation

For every input caption, we ask the LLM to identify a random scene or object that is mentioned in

Model	Dataset	Matching				Retrieval	
		Avg.	Affirmation	Negation	Hybrid	R@5	Neg-R@5
CLIP (Pretrained)	MS-COCO	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)		**53.46**	**68.15**	**45.72**	**45.82**	61.46	59.65
CLIP (Pretrained)	VOC 2007	14.47	31.96	8.34	14.97	N/A	N/A
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)		**61.45**	**86.07**	**38.47**	**75.80**		

Table 1: Results on Negbench MSCOCO and VOC 2007 image datasets on image-to-text matching and text-to-image retrieval tasks. **R@5** refers to the Top-5 recall on original (non-negation) MSCOCO-Caption dataset, while **Neg-R@5** refers to the Top-5 recall on negation MSCOCO-Caption dataset from NegBench.

that caption. The selected scene or object will be used as the negation target to generate the negation caption. Once we have the original caption and the negation target, we generate the negation caption by prompting the LLM to rewrite the original caption such that the negation target should be semantically absent from the original caption.

4.1.2 Evaluation question generation

For each negation caption, we prompt the LLM to identify the positive semantics and the negative semantics in the sentence. For example, given a negation caption *"A dog is playing with a yellow ball while there is no man walking around,"* the positive semantics will be *"A dog is playing with a yellow ball,"* and 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, such as *"Is there a dog playing with a yellow ball?"* and *"Is there a man walking around?"* respectively.

4.1.3 Question quality verification

Although GPT-4o is one of the SOTA LLMs for semantic understanding, it still might generate texts that are semantically incorrect. Thus, verification is necessary to prevent improper generation. Given the negation caption, p, the positive question, q_p, and the negative question q_n, our team members act as the human annotators to manually answer whether the proposition mentioned in the q_p is supported by the caption, (*e.g., "is there a dog?" with the caption "there is a dog but no cat."*), and whether the proposition mentioned in the q_n is contradicted or not supported by the caption, (*e.g., "is there a cat?" with the same caption above*). If the annotators' answer for both questions are correct, the negation data sample will be kept, otherwise it will be discarded. In total, NEG-T2I comprises 2000 human-validated samples, selected from an initial pool of 2500 candidates. The visualization of some samples from the NEG-T2I benchmark is presented in the Appendix A.6.

4.2 Evaluation metrics

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 negation targets. Given an image generated from the negation caption, 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" to the positive question and "no" to the negative question at the same time.

5 Experiments

To show the capability of our proposed method on multiple downstream tasks, we evaluate *TNG-CLIP* 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 using the following benchmarks and the baselines for comparison.

Model	Accuracy
CLIP (Pretrained)	65.16
NegCLIP	73.22
CoN-CLIP	74.15
CLIP (CC12MNegFull)	76.21
NegationCLIP	80.15
TNG-CLIP (Ours)	**81.83**

Table 2: Results on Valse-Existence Benchmark.

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 image-to-text matching task. The benchmark reports the performance as the **accuracy** of correctly matched image-text pairs.
- **NegBench** (Alhamoud et al., 2025b) benchmark is a comprehensive benchmark to evaluate the negation understanding of models on various image-to-text matching and text-to-image retrieval tasks. It includes negation-based matching tasks based on both the MS-COCO (Chen et al., 2015) and the VOC2007 (Everingham et al.) datasets, a negation-based text-to-image retrieval task based on the MS-COCO evaluation dataset. The benchmark reports the matching performance with the **accuracy**, and reports the retrieval performance with the **recall@5**.

Baselines To evaluate the performance of *TNG-CLIP*, we compare it against several existing SOTA 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). All of the methods are initialized based on the pre-trained CLIP ViT-B/32 model.

5.1.1 Comparison experiments

We present the matching and retrieval result of the NegBench-MSCOCO and NegBench-VOC2007 tasks in Table 1. The results indicate that previous methods lack generalization across negation-based tasks and tend to specialize in specific task settings. For example, the *CoN-CLIP*'s accuracy on matching (affirmation) task is 0.05 on the MSCOCO dataset, suggesting the method is biased such that it improves the matching (negation) accuracy at the cost of the affirmation accuracy. For the *NegCLIP* method, even though it gets the best recall on the retrieval task, we observe that the affirmation accuracy is even lower than that of the *pretrained-CLIP*, and its accuracy on matching (hybrid) is low. On the other hand, the *CC12M-NegFull* fine-tuned CLIP presents higher accuracy than all the other baseline methods on different tasks, indicating its capability of diverse negation tasks. Our method, *TNG-CLIP*, while slightly under-performs the *NegCLIP* model on retrieval tasks, achieves the SOTA accuracy on all the matching tasks, showing its generalization and capacity on diverse scenarios.

Similarly, the evaluation on the Valse-Existence dataset, in Table 2, further proves *TNG-CLIP*'s capability of negation understanding. Our method gets the best accuracy, 81.83, higher than all the other negation-understanding CLIP baselines.

5.1.2 Effectiveness of dynamic dataset

Strategy	Avg. Acc.
dynamic dataset	52.09 $\pm$ 1.06
fixed dataset	49.33 $\pm$ 1.34

Table 3: Effect of using dynamic dataset. Evaluation on Negbench-MSCOCO image-to-text matching task.

The training-time data generation pipeline generates the negation caption based on the other image-text pairs in the same batch, which produces different negation captions for the same image in every epoch. We analyze the effect of the dynamic dataset and compare how TNG-CLIP's performance differs from being trained on fixed dataset. We generate the different fixed datasets by storing the image-text sets generated in different training epochs, and use the fixed dataset to fine-tune the CLIP model. To get a statistically significant comparison result, we repeat the *TNG-CLIP*'s training procedure 10 times and use 10 different fixed datasets to fine-tune the pre-trained CLIP.

We present the mean and the standard deviation in Table 3. We observe that the performance of *TNG-CLIP* on MSCOCO matching task is higher than CLIP fine-tuned on the fixed dataset, and the standard deviation is also smaller than the fixed one. We hypothesize that the CLIP's fine-tuning on a fixed dataset constrains the model's negation understanding to specific <*caption, negation object*> pairs, and thus harms the generalization of the model on negation tasks, leading to lower accu-

racy. Also, the data variance among every epoch for *TNG-CLIP* works as a natural regularization to prevent overfitting and memorizing incorrect correlation, thus leading to smaller standard deviation. We also present the ablation study and generation time analysis in the Appendix A.1 and A.2.

5.2 Text-to-image generation evaluation

Model	Arch.	Acc.
SD-1.5	ViT-L/14	44.30
SDXL-1.0	ViT-L/14	40.60
SD-1.5 w/ CoN-CLIP	ViT-L/14	28.55
SD-1.5 w/ TNG-CLIP (ours)	ViT-L/14	**57.35**
pretrained-CLIP + proj	ViT-B/32	29.70
NegCLIP + proj	ViT-B/32	36.90
CoN-CLIP + proj	ViT-B/32	35.55
CC12MNegFull + proj	ViT-B/32	39.90
TNG-CLIP + proj (ours)	ViT-B/32	**48.10**

Table 4: Image Generation on NEG-T2I benchmark

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

5.2.1 Experiment setup

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

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

5.2.2 Experiment analysis

The comparison results on NEG-T2I benchmark are presented in Table 4. The upper table shows the comparison between the CLIP ViT-L/14's text encoders. 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-T2I benchmark. We observe that among the all, using the *TNG-CLIP*'s text encoder achieves the best accuracy, indicating its outstanding capability of handling negation semantics for text-to-image generation. On the other hand, the accuracy of *CoN-CLIP* is lower than the original stable diffusion model, which shows its deficiency on such generation task.

The lower table presents the accuracy of SD-1.5 by replacing its text encoder with the combination of the CLIP's ViT-B/32 based architecture and the fine-tuned MLP projector. We observe that projected ViT-B/32 text encoder is not as effective as ViT-L/14 text encoder, since the accuracy of CLIP's ViT-B/32 text encoder is 29.70, while that for CLIP's ViT-L/14 text encoder is 44.30. However, using *TNG-CLIP*'s text encoder still achieves the highest accuracy, and the clip fine-tuned with *CC12MNegFull* (Alhamoud et al., 2025a) is the second highest, which is similar with its performance in image-text matching tasks. Additional detailed analysis and visualization are in the Appendix A.3.

6 Discussion & Conclusion

In this paper, we focus on the problem of enhancing the 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 captions during the training time, addressing the time- and compute- inefficiency of previous datasets. We also show that using dynamic negation caption during the training can improve model'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, NEG-T2I, to expand the horizon of negation-related benchmarks. Overall, our work underscores the efficiency of negation understanding in the study of vision language models, and calls for the exploration of negation-aware models beyond the matching and retrieval tasks.

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.

Yuliang Cai, Jesse Thomason, and Mohammad Rostami. 2023. Task-attentive transformer architecture for continual learning of vision-and-language tasks using knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6986–7000.

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.

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 experiments

To train the negation-aware CLIP for diverse tasks, we propose a novel asymmetric noise-augmented loss that differs from the CLIP's original contrastive loss. We examine the contribution of each component in this objective function with the ablation study. We decompose our objective function into four components:

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

The analytic results on the NegBench-MSCOCO matching and the NegBench-MSCOCO Retrieval task are presented in Table 5. From the table, we observe that by removing the compositional alignment, the full alignment and the original alignment from the$\mathcal{L}_{t2i}$ separately, the corresponding accuracy in matching task drops. For example, without the original caption, the affirmation accuracy drops from 68.15 to 60.87.

We then analyze the effect of the noises introduced in the $\mathcal{L}_{i2t}$ with five different settings. Instead of choosing a random caption for each image, we match the image to its corresponding (1) original, (2) compositional negation and (3) full negation captions. We also let the image match the (4) random caption chosen from its original, compositional negation and full negation caption. Lastly we (5) remove the $\mathcal{L}_{i2t}$ loss from the objective function. Through the experiments, we found that **without the noise from the randomly paired images and captions, the retrieval task's recall drops significantly.** This supports our hypothesis in Sec 3.2 that the negation dataset is an out-of-distribution task for the pre-trained CLIP, and the direct fine-tuning can degrade the CLIP's performance on negation understanding tasks.

Lastly, we investigate an alternative loss formulation. In the original setup, one image is paired with three captions (C^o, C^c, and C^f), and the contrastive loss is computed jointly over the single image-multiple caption set. In the alternative setup, we instead split this into three separate image-caption pairs and compute the contrastive loss for each pair independently, before summing the results. We present the result at the bottom of Table 5. We observe that the accuracies of both the matching task and the retrieval task drop. Since no random noise is introduced in this formulation, the weaker results further highlight the importance of the noises when fine-tuning CLIP on negation-related datasets.

A.2 Data generation efficiency analysis

In Sec 1, we mention that using LLM-based approach to generate the negation caption is time- and compute-intensive. To explicitly show the necessity of efficient negation caption generation pipeline, we compare the negation caption training dataset generation time for each of the methods (Singh et al., 2024; Park et al., 2025; Alhamoud et al., 2025a). We choose the LLMs for the caption generation based on the choice in their papers, and run the single-sample batch inference using one Nvidia A40 GPU to calculate the approximate generation time for the entire dataset. From Table 7, we observe that *TNG-CLIP* only calls LLM twice to generate the templates for the C^c and the C^f, which is significantly fewer than other methods. Correspondingly, *TNG-CLIP* only takes 7.5 minutes to generate 400K negation captions, while the others takes weeks or months on the single GPU inference setting.

A.3 More Analysis of Image Generation Experiment

In the image generation task, we observe the inefficiency of the original Stable Diffusion and *CoN-CLIP* in the NEG-T2I benchmark. To further explore the reason of such inefficiency, 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 6. In the table, we can observe that for the original Stable Diffusion model, the positive accuracy is higher than that of using our method or *CoN-CLIP* as the 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

Model	Avg.	Affirmation	Negation	Hybrid	Neg-R@5
TNG-CLIP	53.46	68.15	45.72	45.82	59.65
Ablation of Caption Category					
w/o compositional	46.14	62.05	39.09	36.63	55.79
w/o full	51.64	77.95	20.25	54.42	59.84
w/o original	47.49	60.87	46.04	35.33	57.92
Ablation of Noise					
$\mathcal{L}_{i2t}$: original	49.22	66.61	21.89	54.23	43.13
$\mathcal{L}_{i2t}$: compositional	48.02	79.10	13.62	48.61	45.31
$\mathcal{L}_{i2t}$: full	38.12	43.20	41.64	29.72	45.72
$\mathcal{L}_{i2t}$: random of three	46.42	56.09	42.16	40.77	50.01
w/o $\mathcal{L}_{i2t}$	45.11	57.93	35.36	41.71	50.03
independent losses	48.10	62.10	37.11	44.18	49.41

Table 5: Ablation Study on NegBench MSCOCO matching task

Model	Arch.	Positive	Negative
SD-1.5	ViT-L/14	77.80	57.30
SDXL-1.0	ViT-L/14	85.70	47.70
SD-1.5 w/ CoN-CLIP	ViT-L/14	40.40	86.50
SD-1.5 w/ TNG-CLIP (ours)	ViT-L/14	82.95	68.70
pretrained CLIP + proj	ViT-B/32	38.95	78.00
NegCLIP + proj	ViT-B/32	44.95	69.70
CoN-CLIP + proj	ViT-B/32	44.50	83.40
CC12MNegFull + proj	ViT-B/32	53.60	73.95
TNG-CLIP + proj	ViT-B/32	63.20	78.45

Table 6: Image Generation on Neg-T2I benchmark

Model	Training #	LLM inference #	Time
TNG-CLIP	410K	2	7.8 mins
CoN-CLIP	300K	600K	13.8 days
NegationCLIP	229K	605K	12.6 days
CC12MNegFull	10M	30M	173.6 days

Table 7: Single-Sample Batch Negation Caption Generation Time for Different Methods on Single Nvidia A40 GPU

Model	Agreement Rate
SD-1.5	97.20%
SDXL-1.0	96.60%
SD-1.5 w/ CoN-CLIP	95.40%
SD-1.5 w/ TNG-CLIP (ours)	95.60%

Table 8: The accuracy of GPT evaluation

other hand, adopting *CoN-CLIP* as the 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 ignoring the generalization on the other non-negation tasks. We also present the visualization examples in Figure 3. Similarly, we observe that the two original stable diffusion models tend to generate content without the awareness of the negation semantics. *CoN-CLIP*, on the other hand, can successfully avoid the negation objects but also fails to generate the required object.

A.4 Human Assessment of LLM-based Evaluation

In Text-to-image generation task, We deploy the GPT-4o as our evaluator to evaluate the existence and absence of the negation targets. To ensure the reliability of this automatic evaluation approach, we further conduct a human verification study on a randomly selected 500-samples subset of the

Figure 3: Examples of the models' performance on NEG-T2I benchmark.

NEG-T2I. We conduct a human evaluation on the subset's compositional accuracy and compare the results with the GPT-4o evaluation. We computed the agreement rate between the human evaluation result and the GPT evaluation result. We perform such experiments on the ViT-L14 based models and present the results on Table 8.

From the table, we observe that the agreement rates between the human evaluation and GPT evaluation are all above 95%, which indicates that the GPT evaluation pipeline is robust and reliable to evaluate the model's performance on the text-to-image generation task.

A.5 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 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-**

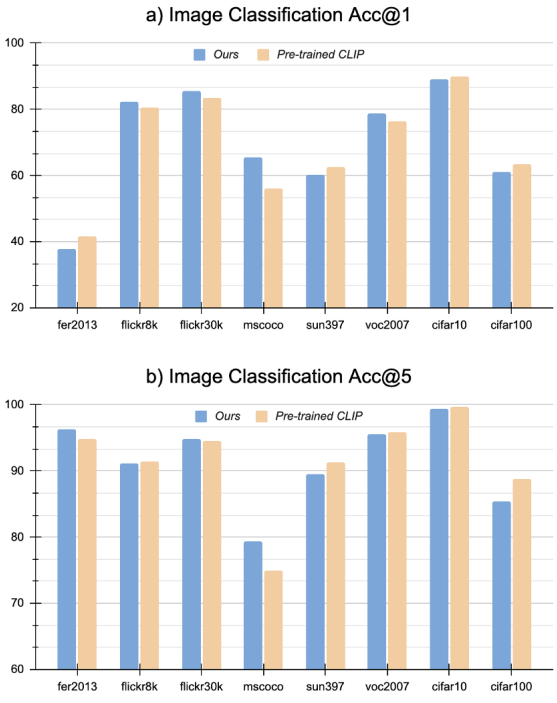

Figure 4: The zero shot image classification accuracy of pre-trained CLIP and TNG-CLIP on eight image classification benchmarks.

10 (Krizhevsky, 2009), **CIFAR-100** (Krizhevsky, 2009). The top1 and top5 accuracy score is presented in Figure 4. 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.6 Negation Pattern & Neg-T2I Visualization

During the negation caption generation, we use predefined LLM-generated negation pattern template to convert the original caption and negation object to compositional negation caption and full negation caption. We present the template we used here in Table 9 and Table 10. We also present the visualization of some data samples from NEG-T2I here in Table 11.

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-T2I* negation image generation benchmark