

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

Junsung Park¹ Jungbeom Lee^{2,3} Jongyoon Song⁴ Sangwon Yu¹ Dahuin Jung^{5†} Sungroh Yoon^{1,6†}

¹Department of Electrical and Computer Engineering, Seoul National University

²Amazon ³Department of Computer Science and Engineering, Korea University

⁴Samsung Research ⁵School of Computer Science and Engineering, Soongsil University

⁶IPAI, AIIS, ASRI, INMC, and ISRC, Seoul National University

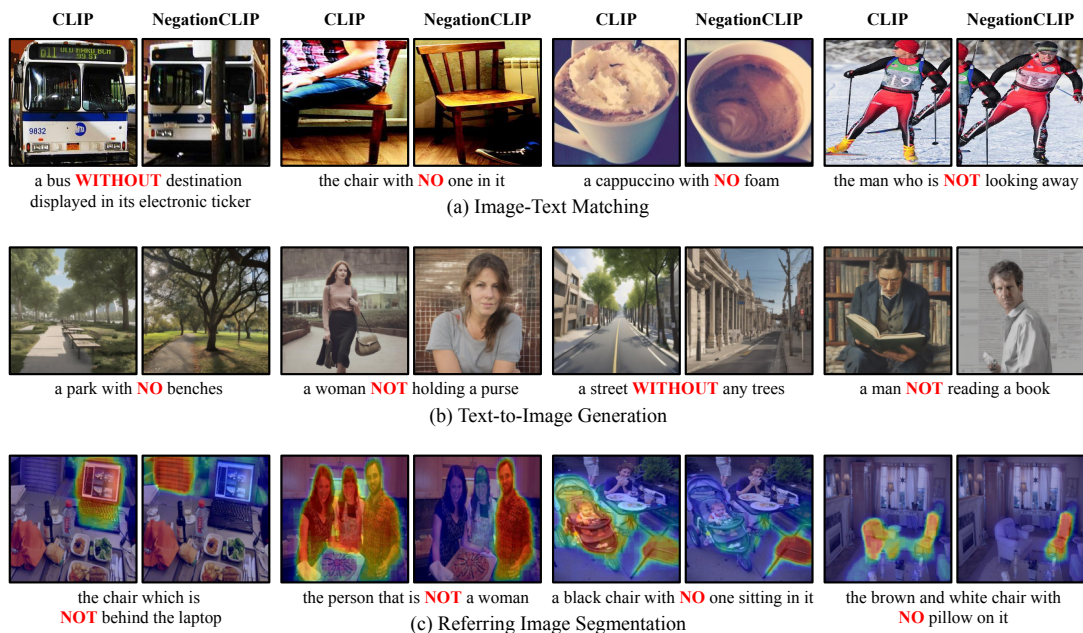


Figure 1. Examples of the original CLIP and NegationCLIP (ours) on negation-inclusive data in multimodal tasks. Our NegationCLIP demonstrates a better understanding of negation concepts across various tasks.

Abstract

While CLIP has significantly advanced multimodal understanding by bridging vision and language, the inability to grasp negation—such as failing to differentiate concepts like “parking” from “no parking”—poses substantial challenges. By analyzing the data used in the public CLIP model’s pre-training, we posit this limitation stems from a lack of negation-inclusive data. To address this, we introduce data generation pipelines that employ a large language model (LLM) and a multimodal LLM to produce negation-inclusive captions. Fine-tuning CLIP with data

generated from our pipelines, we develop NegationCLIP, which enhances negation awareness while preserving the generality. Moreover, to enable a comprehensive evaluation of negation understanding, we propose NegRefCOCOg—a benchmark tailored to test VLMs’ ability to interpret negation across diverse expressions and positions within a sentence. Experiments on various CLIP architectures validate the effectiveness of our data generation pipelines in enhancing CLIP’s ability to perceive negation accurately. Additionally, NegationCLIP’s enhanced negation awareness has practical applications across various multimodal tasks, demonstrated by performance gains in text-to-image generation and referring image segmentation.

[†]Corresponding Authors

1. Introduction

Recent advances in vision-language models (VLMs) [14, 20, 21, 48, 50] have demonstrated significant capabilities in integrating visual and linguistic information, achieving notable performance in multimodal tasks such as text-to-image (T2I) generation [3, 16, 38] and referring image segmentation [17, 18, 28]. Among these models, CLIP [35] has emerged as particularly influential, serving as the foundation for numerous subsequent models [25, 28, 36, 38]. The effectiveness of these models, however, is inherently constrained by the capabilities of the CLIP encoder, underscoring the importance of its robustness.

In response, research has sought to identify and address the inherent limitations of CLIP [5, 8, 42], focusing on challenges related to the text encoder’s ability to understand sentence structure and relationships, which constrain compositional image-text alignment [12, 29, 49]. Despite these efforts, the challenge of accurately handling negation remains largely underexplored, with only a few recent attempts to address this issue [40, 46]. Negation, marked by terms such as “no,” “not,” or “without,” plays a fundamental role in language by altering the meaning of words and phrases and, consequently, entire sentences. Therefore, a precise understanding of negation is crucial for CLIP to perform reliably.

In this study, our preliminary analysis (in Sec. 2) reveals that CLIP frequently fails to capture the intended meaning of prompts involving negation. Experiments underscore this deficiency, demonstrating the need for targeted improvements in this area. Further investigation into the CLIP pre-training dataset [39] indicates that captions containing negation are underrepresented and, when present, often misaligned with the visual content, revealing a critical gap that impedes the model’s ability to understand negation.

To mitigate the limitation in data, we propose two data generation pipelines leveraging a large language model (LLM) and a multimodal LLM (MLLM) to create captions that incorporate negation and align accurately with images. The first pipeline generates negation terms based on the absence of contextually relevant objects, while the second pipeline expands the diversity of negation over object existence. By fine-tuning the pre-trained CLIP text encoder with data generated from our proposed pipelines, we develop NegationCLIP, a model capable of improved comprehension of negation while maintaining general performance across various tasks.

Furthermore, evaluating negation comprehension in VLMs remains challenging due to limited benchmarks and exploration of this issue. Although prior work [29, 33, 40] has introduced image-to-text retrieval benchmarks, they still fall short of providing comprehensive coverage, often exhibiting bias against negation by classifying all negation-inclusive captions as incorrect or restricting negation to

object existence. In light of these limitations, we propose NegRefCOCOg, the first text-to-image retrieval benchmark for evaluating negation comprehension in VLMs. NegRefCOCOg ensures that every retrieval query involves negation and supports a broader range of negation types, applying multiple negation terms not only to objects but also to attributes such as actions, adverbs, and prepositions.

Our experiments demonstrate that NegationCLIP, fine-tuned using data generated by our proposed framework, achieves superior negation understanding on both existing and newly developed NegRefCOCOg benchmarks while maintaining strong general task performance. Additionally, NegationCLIP proves adaptable across various multimodal tasks. Notably, replacing the text encoder in T2I generation models with that of NegationCLIP enhances negation comprehension—a capability often lacking in original T2I models (using the original CLIP encoder). Furthermore, our NegationCLIP’s text encoder enables improved performance in referring image segmentation for prompts containing negation, underscoring its scalability as a more contextually aware and flexible text encoder. Fig. 1 illustrates the versatility of NegationCLIP across diverse multimodal tasks involving negation. The key contributions of our study are summarized as follows:

Contributions 1) We identify a significant limitation in CLIP’s ability to effectively process negation, tracing this issue to deficiencies in the pre-training dataset, where negation terms are underrepresented and poorly aligned with visual content. 2) We develop novel data generation pipelines that utilize a large language model (LLM) and a multimodal LLM (MLLM) to produce high-quality, negation-inclusive captions aligned with visual contexts, enhancing the training data for improved negation comprehension. 3) We propose NegRefCOCOg, a benchmark comprising various forms of negation, specifically designed to evaluate the negation comprehension capabilities of VLMs. 4) Our negation-aware model named NegationCLIP demonstrates robust negation comprehension and maintains generality, excelling across tasks such as image-text matching, T2I generation, and referring image segmentation.

2. Negation: A Critical Challenge for CLIP

In this section, we design a simple experiment to demonstrate that CLIP struggles to handle negation effectively. We then identify potential reasons behind this limitation from a data perspective.

2.1. Case Study: Exposing the Negation Issue

To evaluate CLIP’s ability to handle negation, we conduct a binary classification experiment using the CelebA [26] dataset, which contains 40 binary attributes for facial images. For each of the 40 attributes, we construct positive and

Table 1. Proportion of negation in captions and words within LAION-400M. The table shows the ratio of captions containing negation terms and the ratio of negation terms among all words.

Level	Total Count	Negation Count	Negation Ratio
Caption	414M	2.91M	0.70%
Word	3.88B	3.21M	0.08%

negative prompts following the CLIP-based image classification format: “a photo of.” For example, for the attribute “eyeglasses,” we generate prompts like “a photo of a person wearing glasses” and “a photo of a person *not* wearing glasses”. Detailed prompts are provided in Appendix.

The classification task is structured as follows: given an image, CLIP is prompted with both the positive and negative prompts constructed above, and we evaluate the accuracy in matching the image with the correct prompt out of the two. We use balanced accuracy instead of standard accuracy to account for potential class imbalances between positive and negative examples.

We report the average balanced accuracy of the CLIP ViT-L/14 model across all 40 attributes. Despite this being a binary classification task—where random guessing yields an accuracy of 50%—we obtained an average balanced accuracy of 60.8%. This performance is significantly low, considering that 1,000-way classification with ImageNet-1k produces 73.4% accuracy using the same prompt format. These findings highlight the need for targeted improvements to enhance CLIP’s robustness in handling negation.

2.2. Root Cause: Limited Negation in Training Data

We approach this issue from the perspective of pre-training data. To investigate the presence and quality of negation, we analyze LAION-400M [39], the dataset popularly used for training public CLIP models such as OpenCLIP [2].

In Tab. 1, we present the proportion of captions containing negation in LAION-400M. Our findings reveal that only about 0.704% of captions in LAION-400M contain negation terms. Negation terms make up only 0.083% of the total word count—an insufficient representation given the importance of negation in language. Furthermore, as illustrated in Fig. 2, even when negation is present in captions, it often lacks alignment with the visual content of the image, providing no meaningful signal for the model to learn from. This scarcity of visually aligned negation can be attributed to the nature of image-text pairs typically used in VLM training. Image-level captions naturally focus on describing the contents of an image, detailing what is visible, rather than specifying what is absent. As a result, the model receives minimal exposure to negation-related linguistic structures during training, making it difficult for CLIP to learn and respond to the exclusionary cues introduced by negation.



Figure 2. Examples of misleading negation samples in LAION-400M.

These observations underline a critical need for diverse visually aligned negation-inclusive data, allowing the model to develop a more robust understanding of negation.

3. Negation-Inclusive Data Generation

In this section, we introduce two data generation pipelines using LLM and MLLM to address the limitations in CLIP’s training data—the scarcity of negation terms and their misalignment with visual content. After that, we propose NegRefCOCOg benchmark, which can provide an effective evaluation of negation awareness based on the existing referring image segmentation datasets.

3.1. Generating Negation from Object Absence

In this pipeline, we leverage an existing, well-established image captioning dataset, maximizing utility without the need for extensive new annotations. The goal is to augment existing captions by naturally incorporating negation based on plausible objects that are likely to be present but are actually absent in the image. It has been demonstrated that the absence of objects in an image can be accurately determined by MLLMs [4, 22, 25, 41], making object-based negation a practical and reliable choice for our pipeline.

While the simplest approach in this pipeline might involve a negation regarding random objects, such a method is disconnected from the image context, making it less effective for supporting model learning. Thus, we start by identifying plausible objects using an LLM, to which we provide the caption of the corresponding image. We then use an MLLM to confirm the absence of these plausible objects in the image. For objects confirmed to be absent, we augment the caption using the LLM to naturally incorporate negation terms with their absence, enriching the training data with contextually relevant negation examples. The overall process is as follows:

- 1. Extracting Plausible Object from Caption:** For each image-caption pair, we first provide only the caption to LLM to identify the plausible objects not mentioned in the caption but could reasonably be present in the image.
- 2. Verifying Object Absence with MLLM:** Since the process above does not consider the input image, there is a possibility that the plausible objects identified above

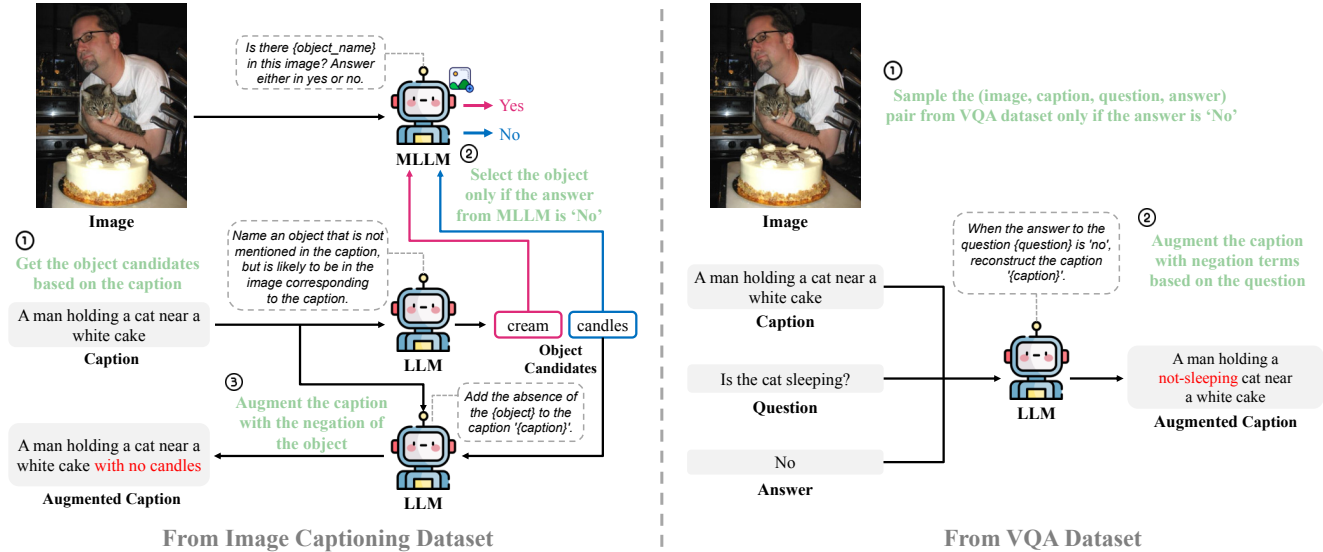


Figure 3. Data generation pipelines in our work. The left presents the pipeline of generating negation from object absence in the image captioning dataset, and the right presents the pipeline of generating negation from question-answer pairs in the VQA dataset.

may actually be present in the image. To confirm their absence, we provide the image to an MLLM and query it about the presence of the identified object. This allows us to filter out objects that are present in the image.

3. **Augmenting Caption with Negation:** For an object that the MLLM determines to be absent, we use an LLM to augment the original caption by adding information about the missing object with negation.

3.2. Expanding Diversity of Negation

The pipeline above is effective but is limited to negations related to objects. To further enrich the data while maximizing the utility of existing resources, we employ a secondary pipeline utilizing data sourced from the VQA dataset. This pipeline introduces diversity in negation expressions by drawing on diverse question-answer pairs, specifically selecting pairs where answers are “no.” These pairs encompass not only object presence but also aspects such as actions being performed and attributes possessed by objects, allowing us to incorporate negation across a broader range of image content. This expansion provides the model with richer exposure to varied linguistic structures. The pipeline operates in the following steps:

1. **Selecting “No” Data:** We identify image-question-answer triplets from the VQA dataset where the answer is “no.” These pairs provide flexibility to incorporate various forms of negation, as they stem from a wide range of questions about different features within the image.
2. **Augmenting Caption with Negation:** For each selected image and its original caption, we use LLM to augment the caption with negation terms based on the question and the corresponding answer.

The overall structure and process of the pipelines are illustrated in Fig. 3, and detailed prompts used for the pipeline are provided in Appendix.

3.3. NegRefCOCOg Benchmark Proposal

While efforts have been made to establish benchmarks for negation evaluation in VLMs, these benchmarks exhibit notable limitations. CREPE Negate [29] and CC-Neg [40] assume that captions containing negation are always false, creating a shortcut where answers can be determined solely by the presence of negation. This severe bias allows even a blind model to outperform VLMs in these benchmarks [12]. VALSE [33], on the other hand, does not impose this bias but remains limited in scope, as it considers only a single negation term (“no”) and restricts negation to object existence. In this work, we address these limitations by proposing NegRefCOCOg, a benchmark built upon the existing referring image segmentation datasets: RefCOCOg [47]. Referring image segmentation datasets can be utilized as valuable sources for negation evaluation because their text prompts contain a relatively higher proportion of negation to distinguish between similar objects within the same image. Additionally, they include diverse negation terms such as “no,” “not,” and “without” and extend negation to various positions within a sentence, covering actions and attributes in addition to object references.

In constructing NegRefCOCOg, we first sample prompts from RefCOCOg that include negation terms. Let T denote a negation-inclusive prompt. For each T , we identify a corresponding image patch P^+ , which aligns with T , serving as the positive example. Additionally, we designate hard negative image patches P^- , representing different instances

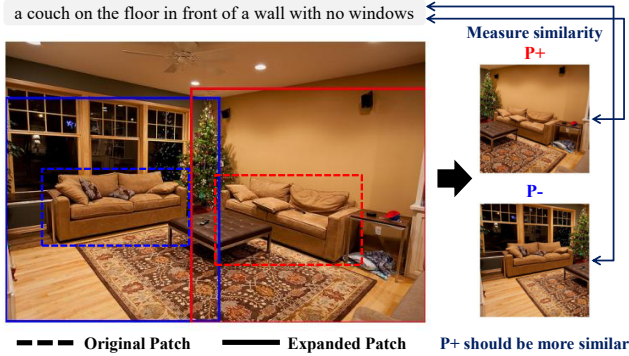


Figure 4. An example of NegRefCOCOg benchmark.

of the same object category with P^+ in a distinct location within the image. This approach ensures that P^- is of the same object type as P^+ but does not align with the negated prompt. After that, we filter and augment the image patches following the constraints that make our benchmark reliable and challenging for evaluating negation understanding.

For evaluation, the model calculates the similarity between the text embedding of T and the vision embeddings of P^+ and P^- . If the similarity between T and P^+ is greater than that between T and P^- , the model scores 1; otherwise, it scores 0. Fig. 4 presents an example of the NegRefCOCOg process. Detailed contents of the constraints and the evaluation can be found in Appendix.

4. Experiments

We evaluate the effectiveness of our negation-inclusive data generation pipeline by fine-tuning the CLIP text encoder and comparing its performance against the original CLIP model and a negation baseline model. Our experiments aim to assess improvements in negation comprehension as well as general performance retention.

4.1. Experimental Setup

Experimental Details We generate a total of 229k image-text pairs using our data generation pipelines, with 147k pairs from the first pipeline (Sec. 3.1) and 82k pairs from the second (Sec. 3.2). For the pipelines, we employ Llama-3-8B [1] as the LLM and LLaVA-1.6 [24] as the MLLM. For datasets, we use COCO [23] as the image captioning dataset and VQAv2 [10] for the VQA dataset.

For fine-tuning, we freeze the vision encoder and fine-tune only the text encoder. This approach helps preserve the original embedding space, making our model, NegationCLIP, adaptable across various tasks without requiring further adjustments. We use the standard InfoNCE loss [31] with a learning rate of $1e-6$, optimized with the AdamW [27] optimizer. Additional prompt design details and specific configurations are provided in Appendix.

Table 2. Comparison of model performance on negation and general benchmarks across different architectures. We include all publicly available CLIP-bnl and CoN-CLIP checkpoints in addition to the baseline CLIP and NegationCLIP (Ours).

Model	Arch.	Negation Benchmarks		General Benchmarks	
		VALSE	NegRefCOCOg	ImageNet	COCO
CLIP		70.97	57.73	62.02	54.78
CLIP-bnl [46]	ViT-B/32	76.78	62.05	53.33	55.47
CoN-CLIP [40]		71.72	55.45	63.08	55.66
NegationCLIP		80.15	64.09	60.97	68.00
CLIP		69.48	58.64	66.71	57.75
CoN-CLIP [40]	ViT-B/16	62.55	55.68	68.57	59.83
NegationCLIP		80.52	64.32	66.33	70.57
CLIP		66.85	57.27	73.44	59.99
CoN-CLIP [40]	ViT-L/14	65.73	55.45	75.38	63.18
NegationCLIP		79.59	62.95	73.91	72.77
CLIP	ViT-L/14 @336px	64.61	57.05	74.92	60.74
NegationCLIP		78.65	62.95	75.15	73.43

Models We evaluate our model, NegationCLIP, with three baselines:

- CLIP [35]: The original, pre-trained CLIP model without any modifications.
- CLIP-bnl [46], CoN-CLIP [40]: Baseline models that have been fine-tuned specifically for negation comprehension, which provides a direct comparison for assessing our approach.
- NegationCLIP: Our proposed model, fine-tuned on negation-inclusive data generated by our two pipelines.

Benchmarks We employ the following benchmarks to assess each model’s performance comprehensively.

- NegRefCOCOg: Our primary benchmark for assessing VLMs’ negation comprehension in an image-text matching task. The model is given a negation-inclusive prompt and two image candidates, and the correct match of prompt and image must be selected.
- VALSE [33] Existence: This benchmark evaluates VLMs’ ability to handle negation in an image-to-text retrieval task. Given an image and two prompts—one indicating an object’s presence and the other its absence—the model must select the prompt that aligns with the image.
- ImageNet [6] classification & COCO [23] retrieval: We evaluate the model’s zero-shot classification accuracy on ImageNet and its recall@5 on COCO text-to-image retrieval, confirming that fine-tuning on negation-inclusive data does not degrade its general capability.

4.2. Main Results

Tab. 2 summarizes the performance of each model across the three evaluation benchmarks.

For **ImageNet** classification and **COCO** retrieval, NegationCLIP maintains accuracy comparable to, or even exceeding, the original CLIP, particularly with larger architectures. This demonstrates that our fine-tuning approach for

Table 3. Ablation results on different data configurations.

Arch.	Data Config.	VALSE \uparrow	NegRefCOCOg \uparrow
ViT-B/32	Original	70.97	57.73
	+ Rand-P1	73.78	62.05
	+ P1	80.15	63.18
	+ P2	76.78	64.32
	+ P1 + P2	80.15	64.09
ViT-B/16	Original	69.48	58.64
	+ Rand-P1	76.22	60.91
	+ P1	77.53	63.41
	+ P2	80.15	63.41
	+ P1 + P2	80.52	64.32
ViT-L/14	Original	66.85	57.27
	+ Rand-P1	76.40	60.23
	+ P1	77.53	60.23
	+ P2	76.03	62.27
	+ P1 + P2	79.59	62.95
ViT-L/14 @336px	Original	64.61	57.05
	+ Rand-P1	76.22	60.00
	+ P1	79.78	60.91
	+ P2	75.28	61.59
	+ P1 + P2	80.34	62.95

negation comprehension does not compromise the model’s general capabilities.

For **NegRefCOCOg** and **VALSE** Existence, Negation-CLIP consistently outperforms the original CLIP and all baseline models across all architectures, demonstrating its effectiveness in interpreting negations. This strong performance can be attributed to how NegationCLIP integrates negation-inclusive captions during training. Unlike CoN-CLIP, which first generates negated captions and then retrieves semantically similar images, our approach begins with an image and generates a corresponding caption that accurately captures its negation-related attributes. By grounding negation-inclusive captions directly in image content, our model learns to better capture the semantic relationships of negated concepts, leading to superior performance in benchmarks requiring fine-grained negation understanding.

Overall, these results confirm that the negation-inclusive data generation pipeline not only significantly improves negation comprehension but also preserves general performance, making NegationCLIP a robust model for handling both negation-specific and broader vision-language tasks.

4.3. Ablation Study on Data Configurations

In this section, we analyze the impact of different data configurations used for fine-tuning to evaluate how effectively they improve the model’s handling of negation. P1 and P2 denote our proposed pipelines described in Sec. 3.1 and Sec. 3.2 respectively. Additionally, we include a Rand-P1

configuration, similar to P1 but using randomly selected objects instead of plausible ones. For more ablation study, please refer to Appendix.

As shown in Tab. 3, across all architectures, the Rand-P1 configuration consistently underperforms compared to P1. This result underscores the importance of plausible object selection in P1, as random object choices lack context, making them less effective for training.

The results indicate that combining P1 and P2 yields the best performance across most negation benchmarks, suggesting that both object-based and VQA-derived negation contribute complementary benefits. While P1 alone performs competitively on the VALSE benchmark, adding P2 further improves performance, especially on NegRefCOCOg. This trend suggests that NegRefCOCOg, in contrast to VALSE, better captures the diversity of negation expressions, including those based on actions and attributes. Consequently, the combined configuration (P1 + P2) aligns well with NegRefCOCOg’s expanded scope, leading to superior performance in negation understanding.

5. Application

To further validate our negation-inclusive fine-tuning approach, we apply the model across different multimodal tasks. Specifically, we evaluate its performance on T2I generation and referring image segmentation, demonstrating improvements in negation comprehension.

5.1. Text-to-Image Generation with Negation

T2I models often rely on the CLIP text encoder for interpreting prompts but struggle with negation, a limitation that aligns with CLIP’s challenges in handling negation. For instance, given a prompt like “a man not wearing a hat,” a T2I model using the original CLIP text encoder may generate an image of a man wearing a hat, ignoring the negation.

To evaluate our model’s ability to handle negation in T2I tasks, we replace the original CLIP text encoders in Stable Diffusion [38] models with our NegationCLIP text encoder, without further training on the T2I model. This direct substitution is possible because we fine-tune only the text encoder, preserving the original image embedding space and maintaining alignment with it.

We utilize ChatGPT [32] to generate a set of 107 negation-inclusive prompts to measure each model’s ability to represent negated concepts in generated images. We assess the performance of negation comprehension using an MLLM evaluator, mPLUG [19], inspired by the evaluation methodology employed in TIFA [13]. We provide the MLLM with the generated image and two queries: (1) whether the subject is present in the image, and (2) whether the negation-related concept is correctly excluded. If both conditions are met, we assign a score of 1; otherwise, a



Figure 5. Examples of text-to-image generation on negation-inclusive prompts.

Table 4. Comparison of text-to-image generation performance on negation-inclusive prompts.

Model	TIFA \uparrow	Neg Score \uparrow
SD-1.4 [37]	0.786	0.295
SD-1.4 w/ CoN-CLIP text encoder	0.783	0.243
SD-1.4 w/ NegationCLIP text encoder	0.790	0.449
SDXL-1.0 [34]	0.849	0.308
SDXL-1.0 w/ NegationCLIP text encoder	0.802	0.421

score of 0. For details on the specific prompts used for image generation and MLLM evaluation, please refer to Appendix.

For each of the 107 prompts, we generate images using 5 different random seeds, and the average score across these generations is reported as the **Neg Score** in Tab. 4. Our model achieves significantly higher Neg Scores across both SD-1.4 [37] and SDXL-1.0 [34], with over 0.15 improvement compared to the original text encoder, indicating enhanced negation comprehension, while CoN-CLIP’s low Neg Score further confirms the issues arising from the misalignment between its negation-inclusive captions and retrieved images, as observed in retrieval benchmarks. Qualitatively, as shown in Fig. 5, our model successfully generates images reflecting prompts like “a dog not running” or “a street with no lights,” whereas the original model often fails to capture the negation.

To ensure that negation-aware fine-tuning of CLIP does not compromise general T2I quality, we also evaluate each model on the TIFA [13] benchmark, which assesses general text-image alignment. Tab. 4 shows that our model retains competitive TIFA scores, slightly exceeding the orig-

Table 5. Comparison of referring image segmentation performance on PhraseCut and RefCOCOg (Neg).

Model	PhraseCut		RefCOCOg (Neg)	
	mIoU	IoU _{BIN}	mIoU	IoU _{BIN}
CLIPSeg [28]	0.562	0.736	0.267	0.492
CoN-CLIPSeg	0.539	0.724	0.123	0.379
NegationCLIPSeg	0.561	0.737	0.288	0.521

inal SD-1.4 model and trailing slightly in SDXL-1.0. The slight decrease in SDXL’s performance may stem from its use of dual text encoders, both of which were replaced with our negation-aware encoder. This setup could introduce minor alignment challenges, which might be further improved through the additional adaption of diffusion models to our NegationCLIP text encoder.

Overall, our approach addresses a key limitation in NegationCLIP-based T2I models, enabling more accurate generation in negation contexts.

5.2. Referring Image Segmentation

Referring image segmentation task requires models to segment regions within an image corresponding to the given text prompts. To assess the effectiveness of our NegationCLIP, we replace the text encoder in the existing referring image segmentation model, CLIPSeg [28], with NegationCLIP text encoder. We simply replace the text encoder without further training as we did in Sec. 5.1. We refer to this model as NegationCLIPSeg and the same goes for CoN-CLIPSeg.

Tab. 5 presents the mIoU and IOU_{BIN} as done in CLIPSeg [28]. On the PhraseCut dataset, which lacks nega-

tion in prompts, the performance of NegationCLIPSeg is comparable to the original model. On the negation-inclusive subset of RefCOCOg, NegationCLIPSeg demonstrates a performance boost, achieving higher mIoU and IoU_{BIN} scores compared to the original CLIPSeg. This improvement highlights NegationCLIPSeg’s enhanced capability to handle prompts with negation, while still retaining general performance.

Fig. 6 shows qualitative examples from the RefCOCOg negation subset. In each example, NegationCLIPSeg generates more accurate segmentations that align with the negation-specific prompts, whereas the original CLIPSeg model often fails to correctly interpret the negation. For instance, given the prompt “a man with a beard *not* riding an elephant,” NegationCLIPSeg successfully excludes the people riding an elephant, focusing mostly on the man not riding an elephant.

Overall, these results highlight the potential of incorporating a negation-aware text encoder to improve performance in referring image segmentation tasks in scenarios that require precise comprehension of negation within multimodal contexts.

6. Related Work

Vision-Language Models and Their Limitations VLMs [14, 35, 48] learn joint image-text embeddings from large-scale paired data, enabling diverse multimodal tasks such as image classification and zero-shot retrieval.

However, numerous studies have highlighted limitations in VLMs, particularly concerning the models’ handling of compositional language, object relationships, and fine-grained language distinctions [12, 29, 49]. To address such limitations, researchers have explored targeted data generation techniques and fine-tuning methods. Data augmentation approaches [43, 45] aim to improve the robustness of VLMs by generating or augmenting data with challenging linguistic structures. Despite these improvements, the issue of negation remains relatively underexplored.

Understanding and Addressing Negation Negation has been a challenging aspect in language models, as it plays a crucial role in altering or reversing the meaning of a phrase. While some research has addressed negation comprehension in natural language processing (NLP), such as studies highlighting significant issues in BERT’s [7] interpretation of negated statements [9, 11, 15], there has been relatively little exploration of negation within multimodal models. In the vision-language domain, only a handful of studies have specifically focused on incorporating negation, and even these efforts have limited scope.

Existing studies have integrated negation into CLIP for out-of-distribution (OoD) detection tasks [30, 44], while others have focused on specific applications, such as nega-

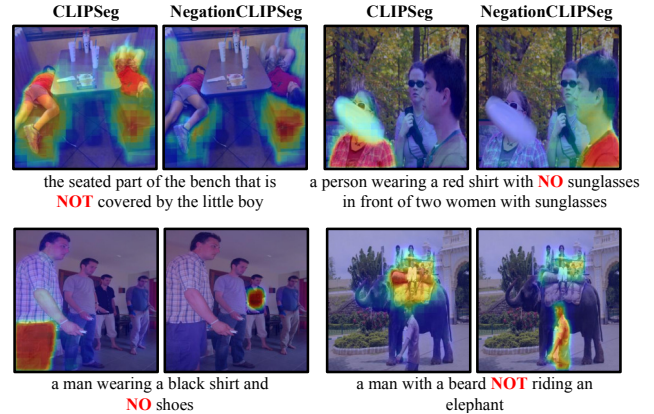


Figure 6. Examples of referring image segmentation on negation-inclusive prompts.

tion handling in video tasks [46]. However, these models are restricted to the specific contexts, making these approaches less adaptable for broader downstream tasks. A more recent approach [40] seeks to enhance CLIP’s ability to process negation beyond task-specific settings through fine-tuning, using image-caption pairs where negation-inclusive captions are generated first and then matched with similar images retrieved from a fixed pool.

Our work extends this line of research by developing a negation-inclusive text encoder within CLIP, designed for flexible, plug-and-play integration across a variety of multimodal tasks, addressing the underexplored challenge of negation comprehension in VLMs.

7. Conclusion

This study addresses a critical limitation in CLIP by introducing a negation-inclusive fine-tuning approach that significantly enhances the models’ ability to interpret negated prompts. Our data generation pipeline leverages large language models to create diverse negation-inclusive captions, enabling fine-tuning that effectively bridges the gap in negation comprehension. Experimental results on the VALSE benchmark and our proposed NegRefCOCOg benchmark demonstrate substantial improvements in our NegationCLIP over the original CLIP in handling negation, while maintaining strong performance on general benchmarks. Furthermore, we show that NegationCLIP enables effective applications of negation across various multimodal tasks, including text-to-image generation and referring image segmentation.

Our work underscores the potential of targeted data generation in advancing the semantic capabilities of CLIP. By refining their ability to process negation, we move closer to developing a VLM that is more aligned with the subtleties of human language, ultimately making them better suited for complex multimodal tasks.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A3B1077720, 2022R1A5A7083908, RS-2025-00555943), Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [NO. RS-2021-II211343, RS-2022-II220959, Artificial Intelligence Graduate School Program (Seoul National University), No.RS-2025-02263754, Human-Centric Embodied AI Agents with Autonomous Decision-Making], the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2025, and Samsung Electronics (IO221213-04119-01).

References

- [1] AI@Meta. Llama 3 model card. 2024. 5
- [2] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 3
- [3] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11472–11481, 2022. 2
- [4] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 3
- [5] Sepehr Dehdashtian, Lan Wang, and Vishnu Naresh Boddeti. Fairerclip: Debiasing clip’s zero-shot predictions using functions in rkhs. *arXiv preprint arXiv:2403.15593*, 2024. 2
- [6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5
- [7] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 8
- [8] Lei Fan, Jianxiong Zhou, Xiaoying Xing, and Ying Wu. Active open-vocabulary recognition: Let intelligent moving mitigate clip limitations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16394–16403, 2024. 2
- [9] Iker García-Ferrero, Begoña Altuna, Javier Álvez, Itziar Gonzalez-Dios, and German Rigau. This is not a dataset: A large negation benchmark to challenge large language models. *arXiv preprint arXiv:2310.15941*, 2023. 8
- [10] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 5
- [11] Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordani, and Aaron Courville. Understanding by understanding not: Modeling negation in language models. *arXiv preprint arXiv:2105.03519*, 2021. 8
- [12] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023. 2, 4, 8
- [13] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023. 6, 7
- [14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2, 8
- [15] Aditya Khandelwal and Suraj Sawant. Negbert: a transfer learning approach for negation detection and scope resolution. *arXiv preprint arXiv:1911.04211*, 2019. 8
- [16] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 2
- [17] Jungbeom Lee, Sungjin Lee, Jinseok Nam, Seunghak Yu, Jaeyoung Do, and Tara Taghavi. Weakly supervised referring image segmentation with intra-chunk and inter-chunk consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21870–21881, 2023. 2
- [18] Jungbeom Lee, Sanghyuk Chun, and Sangdoon Yun. Toward interactive regional understanding in vision-large language models. *arXiv preprint arXiv:2403.18260*, 2024. 2
- [19] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022. 6
- [20] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2
- [21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2
- [22] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 3
- [23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence

- Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5
- [24] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 5
- [25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 3
- [26] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 2
- [27] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5
- [28] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7086–7096, 2022. 2, 7
- [29] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921, 2023. 2, 4, 8
- [30] Jun Nie, Yonggang Zhang, Zhen Fang, Tongliang Liu, Bo Han, and Xinmei Tian. Out-of-distribution detection with negative prompts. In *The Twelfth International Conference on Learning Representations*, 2024. 8
- [31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5
- [32] OpenAI. Gpt-4 technical report, 2023. 6
- [33] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*, 2021. 2, 4, 5
- [34] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 7
- [35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 5, 8
- [36] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 2
- [37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 7
- [38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 6
- [39] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2, 3
- [40] Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. Learn” no” to say” yes” better: Improving vision-language models via negations. *arXiv preprint arXiv:2403.20312*, 2024. 2, 4, 5, 8
- [41] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 3
- [42] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13019–13029, 2024. 2
- [43] Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26700–26709, 2024. 8
- [44] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1802–1812, 2023. 8
- [45] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907*, 2023. 8
- [46] Ziyue Wang, Aozhu Chen, Fan Hu, and Xirong Li. Learn to understand negation in video retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 434–443, 2022. 2, 5, 8
- [47] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 4
- [48] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2, 8

- [49] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022. [2](#), [8](#)
- [50] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. [2](#)