Targeted Caption Generation Improves Compositional Reasoning in VLMs

Arjun Chandra Patrick Lutz Advik Vyas
Boston University
{ac25, plutz, vadvik}@bu.edu

Abstract

Vision-language models (VLMs) have achieved remarkable success in zero-shot recognition tasks, yet their ability to perform compositional reasoning—understanding relationships between objects, attributes, and spatial configurations—remains limited. While various fine-tuning methods have been proposed to address this shortcoming, we find that performance gains in current state-of-the-art approaches often arise from exploiting simplistic negative captions rather than fostering genuine compositional reasoning. We introduce High-quality Targeted Caption Generation, a fine-tuning framework that employs semantically coherent positive and negative captions to instill true compositional reasoning abilities in VLMs. We demonstrate that our method improves performance on challenging compositional reasoning benchmarks, highlighting the critical role of caption quality and offering a promising path toward more robust, reasoning-capable VLMs.

1. Introduction

Vision-language models (VLMs) such as CLIP and BLIP [9, 15] have revolutionized multimodal AI by aligning visual and textual representations in a shared embedding space. These models excel at a range of vision-language tasks, including object recognition and visual question answering. However, they still struggle with compositional reasoning—the ability to understand complex relationships between objects, attributes, and actions. This limitation poses challenges for real-world applications requiring fine-grained visio-linguistic understanding, such as autonomous systems and assistive technologies.

In this work, we analyze state-of-the-art fine-tuning strategies aimed at improving compositional reasoning in VLMs. Specifically, we focus our efforts on the recently proposed DAC-SAM [3], as this method yields the best performance in average across 12 compositional reasoning benchmarks [13]. Our analysis, however, reveals that DAC-SAM's rule-based caption negation used for fine-tuning encourages the model to simply identify nonsensical captions as negatives—leading to inflated performance on easily "hackable" benchmarks such as CREPE and ARO [5]. To this end, we propose **H**igh-quality **T**argeted **C**aption **G**eneration (HTCG), a fine-tuning framework that leverages visually-aligned positive captions and coherent, but compositionally distinct negative captions. We summarize our main findings:

- Analogy-style evaluations in a VLM's embedding space do not reliably measure compositional reasoning ability.
- DAC-SAM overfits to its rule-based caption negation which often results in incoherent negative captions, limiting its ability to generalize to challenging compositional benchmarks.
- High-quality targeted caption generation promotes robust compositional reasoning in VLMs.

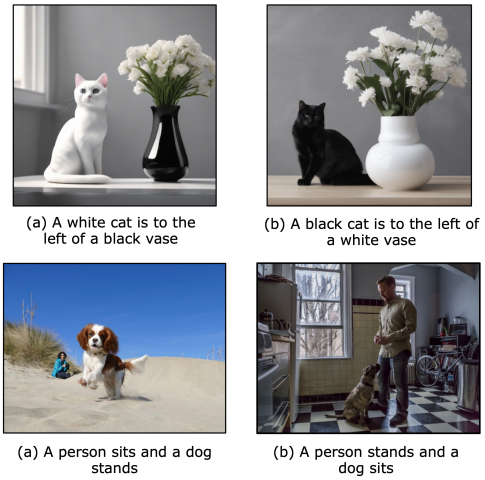

Figure 1. Examples from compositional reasoning benchmarks.

2. Related Work

VLMs and Compositional Reasoning. Vision-language models (VLMs) such as CLIP [15] have achieved impressive results by aligning visual and textual modalities in a shared embedding space using contrastive learning on web-

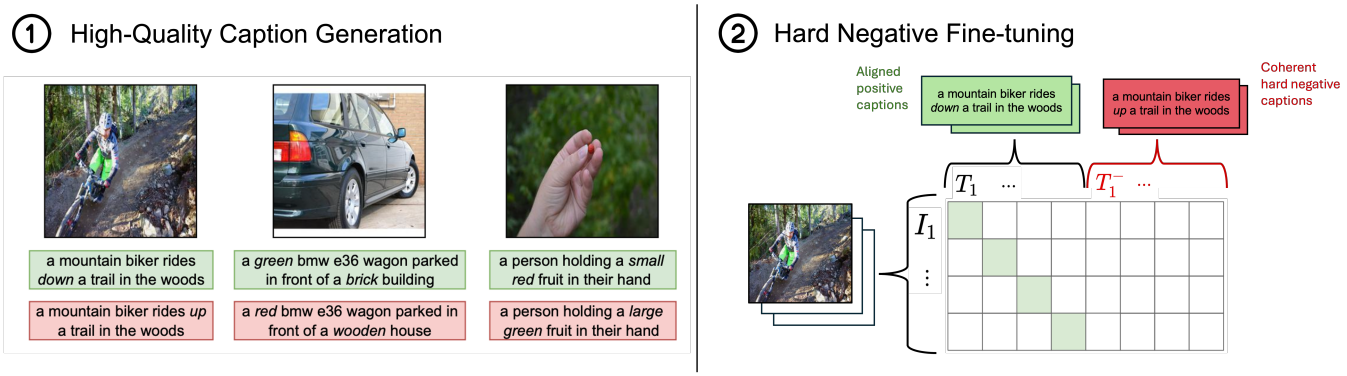

Figure 2. Overview of our proposed approach. We first generate high-quality captions for 1.7M images from the Open Images dataset [8]. Positive captions are generated using BLIP-2-6.7b [10] to ensure alignment with the visual content (Section 3.3.1), while negative captions are generated by Mistral-7B-Instruct-v0.3 [6] to be compositionally distinct from the positive caption while remaining semantically coherent (Section 3.3.2). The generated captions are then used to fine-tune the base ViT-B-32 model [15] (Section 3.3.3).

scale datasets. More recent models, like BLIP [9], introduce task-specific objectives and cross-attention mechanisms to improve modality interaction. Despite these advances, benchmarks such as Winoground [17], SugarCrepe [5], and ColorSwap [1] consistently show that pre-trained VLMs struggle with compositional reasoning, often behaving more akin to bags-of-words models [18].

Improving Compositional Reasoning. Numerous approaches have been proposed to improve compositional reasoning in VLMs, which can be broadly categorized into pre-training, fine-tuning, and training-free methods. Our work falls under the fine-tuning category. Early fine-tuning approaches such as NegCLIP and TSVLC [4, 18] generate hard-negative captions using rule-based augmentations. Subsequent methods like CLoVe and DAC-SAM [2, 3] also generate positive captions for each image to ensure alignment with its visual content. DAC-SAM further incorporates Multiple Instance Learning (MIL) to leverage several positive and negative captions for each image. More recent techniques, such as GNM-CLIP [16], introduce negative images during fine-tuning to provide stronger supervision in the visual space. Similar to these works, our method generates both positive and negative captions for each image; however, we employ a large language model (LLM) to generate semantically coherent and targeted negative captions.

3. Methods

3.1. Effect of incoherent negative captions

Among existing fine-tuning methods, DAC-SAM achieves the highest average performance across the 12 benchmarks evaluated in Oh et al. [13]. However, these gains are disproportionately concentrated on benchmarks susceptible to superficial cues—e.g., a >25% improvement on CREPE—while performance on more robust benchmarks such as Winoground remains unchanged (i.e., +0%). We hypothesize that this discrepancy arises from the rule-based caption negator used during DAC-SAM's fine-tuning, which frequently generates grammatically incorrect or semantically incoherent captions. This study examines that hypothesis and quantifies the negator's influence on model behavior.

The DAC-SAM negator generates negative captions by altering object attributes, relations, or states in the original descriptions, aiming to promote compositional reasoning. Yet prior work [5] suggests these synthetic negatives often reduce to linguistic shortcuts, producing implausible captions that models can reject without true semantic understanding.

To evaluate this, we apply DAC-SAM's negator to generate alternative negative captions for existing benchmarks (we drop the original negative captions) and report the performance gain from the alternative negative captions compared to the original captions (Table 1). First, the performance gain for the base ViT-B-32 model reflects the extent to which each benchmark becomes easier when using the negator-generated negative captions. We find that, across all benchmarks, the negator consistently produces lower-quality captions than the original ones found in the benchmarks. Second, we define Δ as the difference between the ViT-B-32 gain and the DAC-SAM gain. A large Δ indicates that DAC-SAM had already captured most of the performance gains associated with exploiting low-quality captions present in the original benchmark—particularly in CREPE and ARO. Conversely, negative Δ values on ColorSwap and Winoground suggest that DAC-SAM fails to identify incoherent captions in these higher-quality benchmarks, only improving when such captions are artificially

introduced through our intervention.

In sum, our findings indicate that DAC-SAM's fine-tuning process, driven by a flawed negator, biases the model toward detecting incoherent captions rather than learning genuine compositional reasoning. This yields strong results on benchmarks susceptible to such shortcuts, but limited gains on more rigorous evaluations.

3.2. Probing embedding quality

To further explore whether DAC-SAM exhibits true compositional understanding, we follow the analysis from [7] and evaluate whether the text and image encoders exhibit linear analogies in their embedding space. Specifically, we select 13 objects which are placed around a table in the What'sUp benchmark [7]. We then evaluate, for example, whether $I(\text{banana on table}) - I(\text{banana under table}) + I(\text{cup under table})$ is the closest to $I(\text{cup on table})$, compared to $I(\text{cup left/right/under table})$, where $I(\cdot)$ is the text/image encoder. Given 13 objects and 4 prepositions, there are 468 such analogies, and we report accuracy as the percentage of analogies where the condition holds. For reference, we also evaluate the base ViT-B-32 and BLIP. Results are shown in Table 2.

While all models exhibit a performance gap between the text and image embedding spaces, DAC-SAM shows the most noticeable divergence. Its text encoder achieves near-perfect analogy accuracy (99.8%), a substantial improvement over ViT-B-32 (90.6%) and BLIP (80.9%). However, its image encoder collapses to just 1.3% accuracy compared to ViT-B-32's 7.2% and BLIP's 5.9%. This result suggests that DAC-SAM's fine-tuning process strongly overfits to textual patterns, likely due to rule-based caption negation, while severely degrading visual grounding. These findings are consistent with our observations that DAC-SAM's improvements are concentrated on benchmarks with hackable linguistic shortcuts.

However, we are careful not to over-interpret embedding analogy results as definitive evidence of compositional understanding. For instance, BLIP underperforms on this analogy task despite greatly outperforming ViT-B-32 (+75%) on harder compositional benchmarks like ColorSwap [1]. This discrepancy highlights a limitation of this analysis as it assumes semantic and compositional relationships are captured linearly in the embedding space, which may not be a well-founded assumption. Another limitation of this analysis is that contrasting images in the What'sUp dataset are often globally very similar. As a result, a common failure case arises when, for example, $I(\text{banana on table}) - I(\text{banana under table}) \approx 0$, since only a small local feature—namely, the position of the banana—differs between the two images. This minimal visual difference leads to nearly identical embeddings, and we suspect this to be the primary reason why all three models perform significantly worse in the image space on this task.

	Benchmark				
Model	CREPE	ARO	ColorSwap	SugarCrepe	Winoground
BLIP	+4.8	+10.0	-12.3	-15.3	+1.1
ViT-B-32	+16.7	+29.1	+24.1	+4.8	+24.0
DAC-SAM	+3.3	+3.3	+38.4	+4.2	+60.7
Δ	+13.4	+9.5	-14.3	+0.6	-36.7

Table 1. One-way text score gain (%) with DAC-SAM negator over ground-truth negative caption. Δ indicates the difference in gain between ViT-B-32 and DAC-SAM.

Model	Text	Image
ViT-B-32	0.906	**0.072**
BLIP	0.809	0.059
DAC-SAM	**0.998**	0.013

Table 2. Accuracy on compositional embedding analogies in text and image space

3.3. TCG: Targeted Caption Generation

To address the limitations of previous approaches, we construct a new fine-tuning dataset using 1.7M images from the Open Images V7 training split [8]. For each image, we generate one positive caption that is closely aligned with its visual content, and one negative caption that introduces a minimal compositional contradiction. Examples from our fine-tuning dataset are shown in Figure 2. To train the model to distinguish between these paired captions, we modify the standard InfoNCE loss for hard negative fine-tuning [18].

3.3.1. Image-Aligned Positive Captions

Following previous works [2, 3], we note that captions in web-scraped datasets are often noisy or weakly aligned with the image. Therefore, the goal of generating positive captions is to provide accurate, visually grounded descriptions that reflect the content of the image. To achieve this, we caption each image using BLIP-2-6.7b [10]. Unlike DAC-SAM, we only generate one positive caption per image.

3.3.2. Plausible Hard Negative Captions

The goal of the negative captions is to be linguistically coherent while remaining compositionally distinct from the image and its corresponding positive caption. To achieve this, similar to [14], we prompt Mistral-7B-Instruct-v0.3 [6] with the positive caption and instruct it to generate a minimally modified alternative that could not possibly describe the same image (i.e., by swapping objects, attributes, or relations). We use a small number of few-shot examples to guide the generation process and encourage consistency. Leveraging a powerful LLM ensures that the generated hard negatives are both compositionally distinct and

Model	Challenging Benchmarks			Hackable Benchmarks	
	ColorSwap	SugarCrepe	Winoground	ARO	CREPE
ViT-B-32‡[15]	0.137	0.765	0.080	0.588	0.648
NegCLIP† [18]	0.183	0.837	0.080	0.801	0.303
CLoVe† [2]	**0.186**	0.845	0.065	0.829	0.416
CE-CLIP† [19]	0.133	0.856	0.0525	0.763	0.346
GNM-CLIP† [16]	0.126	0.787	**0.103**	0.571	0.173
TSVLC† [4]	0.107	0.769	0.0675	**0.835**	0.359
DAC-SAM‡ [3]	0.122	0.866	0.080	0.725	**0.902**
HTCG (ours)	0.159	**0.897**	0.070	0.666	0.777

Table 3. Comparison of fine-tuning methods on challenging benchmarks and hackable benchmarks. All methods are applied to the pre-trained OpenAI CLIP ViT-B-32 [15]. For ARO, CREPE, and SugarCrepe we report text score. For ColorSwap and Winoground we report group score. †Numbers taken from [13] ‡Our implementation.

semantically meaningful, mitigating the grammatical artifacts that can make synthetic negatives easy to detect.

3.3.3. Hard Negative Fine-tuning

To incorporate hard negative captions during fine-tuning, we modify the standard InfoNCE loss by extending the image-text contrastive term to include the negative captions:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\mathbf{z}_i^{\text{img}}\cdot\mathbf{z}_i^{\text{text}}/\tau\right)}{\sum_{k=1}^{\mathbf{2N}}\exp\left(\mathbf{z}_i^{\text{img}}\cdot\mathbf{z}_k^{\text{text}}/\tau\right)} \tag{1}$$
$$-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\mathbf{z}_i^{\text{text}}\cdot\mathbf{z}_i^{\text{img}}/\tau\right)}{\sum_{k=1}^{N}\exp\left(\mathbf{z}_i^{\text{text}}\cdot\mathbf{z}_k^{\text{img}}/\tau\right)}$$

Indeed, this encourages the model to distinguish the correct caption not only from other positive samples in the batch, but also from semantically and compositionally similar yet incorrect (i.e., hard negative) captions.

4. Experiments

Training. To evaluate our approach, we fine-tune the base ViT-B-32 model on our constructed dataset for 3 epochs with a batch size of 32 using the AdamW optimizer. The maximum learning rate is set to 5e-4, and we use a linear warmup over the first 2,000 steps followed by cosine decay. To preserve the general utility of the base model, the pre-trained model weights are kept frozen, and we apply LoRA adapters ($r = 4$, $\alpha = 8$) to each attention layer in both the text and vision towers. Training takes approximately 4 hours on a single A100 GPU.

Benchmarks and Baselines. We evaluate the base model, our fine-tuning method, and six previously proposed fine-tuning methods across five benchmarks: CREPE [12] and ARO [18], classified as hackable by Hsieh et al. [5], and ColorSwap [1], Winoground [17], and SugarCrepe [5], generally considered more challenging and robust.

Results & Analysis. Table 3 reports the performance of the base ViT-B-32 model alongside various fine-tuning methods across multiple benchmarks.

First, our model improves upon the base model on all benchmarks except Winoground, where no fine-tuning method yields significant gains. This suggests that our approach effectively enhances the base model's compositional reasoning capabilities.

Second, our model outperforms DAC-SAM—the current state-of-the-art in average performance across compositional benchmarks [13]—on two challenging datasets, ColorSwap and Winoground, while underperforming on more easily exploitable benchmarks. As discussed in Section 3.1, this is likely because DAC-SAM relies on detecting incoherent captions rather than true multimodal reasoning.

Finally, our model sets a new state-of-the-art on SugarCrepe, highlighting the effectiveness of our fine-tuning approach relative to existing methods. Notably, both NegCLIP and CE-CLIP are fine-tuned on the COCO dataset, from which SugarCrepe's data is curated [5, 11]. In contrast, our model is trained without any exposure to COCO, thus operating under a greater domain shift—yet still achieving superior performance.

Limitations. We identify failure cases in our LLM-based caption generation. In some instances, the negative caption does not meaningfully differ from the positive one (e.g., "the ocean and beach in Miami" vs. "the beach and ocean in Miami"), while in others, the semantic shift is too large to constitute a hard negative (e.g., "a young boy in a green hoodie and black pants standing on a sidewalk" vs. "an old man in a red hoodie and white pants sitting on a grassy

hill"). These issues could likely be mitigated through improved prompting strategies or by employing more capable language models.

Additionally, while all fine-tuning methods are applied to the same base model to ensure comparability, they rely on different fine-tuning datasets that vary in both size and origin. Understanding how performance is affected by the choice of fine-tuning data would be highly informative. For instance, it remains an open question to what extent our performance gains on SugarCrepe might be driven by our choice of fine-tuning data.

5. Conclusion

In conclusion, our work revisits the challenge of improving compositional reasoning in vision-language models. We show that current state-of-the-art methods like DAC-SAM achieve strong benchmark performance by exploiting grammatical irregularities in negative captions, rather than by learning genuine compositional structure. To address this, we propose High-quality Targeted Caption Generation (HTCG), a fine-tuning framework that pairs visually aligned positive captions from a captioning model with high-quality, semantically coherent negative captions generated by a large language model. This approach is general and can be applied to any image dataset. Our experiments show that HTCG enhances compositional reasoning, particularly on challenging and less hackable benchmarks. These results underscore the importance of caption quality in fine-tuning and suggest that careful data curation is crucial to instilling genuine reasoning capabilities in vision-language models.

6. Team Member Contributions

All team members contributed to writing the report. Additional team member contributions are as follows:

1. **Arjun** – Conducted embedding analogies experiment (Section 3.2), generated negative captions (Section 3.3.2), and implemented fine-tuning (Section 4). Explored style-transfer as a strategy to address distribution shifts in synthetic data (direction dropped).
2. **Patrick** – Conducted DAC-SAM negator experiment (Section 3.1), generated positive captions (Section 3.3.1), constructed the codebase, and benchmarked the models (Table 3).
3. **Advik** – Benchmarked ViT-B-32 on What'sUp prepositions.

References

[1] Jirayu Burapacheep, Ishan Gaur, Agam Bhatia, and Tristan Thrush. Colorswap: A color and word order dataset for multimodal evaluation, 2024. 2, 3, 4

[2] Santiago Castro, Amir Ziai, Avneesh Saluja, Zhuoning Yuan, and Rada Mihalcea. Clove: Encoding compositional language in contrastive vision-language models, 2024. 2, 3, 4

[3] Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. Dense and aligned captions (dac) promote compositional reasoning in vl models, 2023. 1, 2, 3, 4

[4] Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. Teaching structured visionlanguage concepts to visionlanguage models, 2023. 2, 4

[5] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality, 2023. 1, 2, 4

[6] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. 2, 3

[7] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning, 2023. 3

[8] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 2, 3

[9] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 1, 2

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 2, 3

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 4

[12] 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. 4

[13] Youngtaek Oh, Pyunghwan Ahn, Jinhyung Kim, Gwangmo Song, Soonyoung Lee, In So Kweon, and Junmo Kim. Exploring the spectrum of visio-linguistic compositionality and recognition, 2024. 1, 2, 4

[14] Maitreya Patel, Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, and Yezhou Yang. Tripletclip: Improving compositional reasoning of clip via synthetic vision-language negatives, 2024. 3

[15] 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. Learning transferable visual models from natural language supervision, 2021. 1, 2, 4

[16] Ugur Sahin, Hang Li, Qadeer Khan, Daniel Cremers, and Volker Tresp. Enhancing multimodal compositional reasoning of visual language models with generative negative mining, 2023. 2, 4

[17] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality, 2022. 2, 4

[18] 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?, 2023. 2, 3, 4

[19] Le Zhang, Rabiul Awal, and Aishwarya Agrawal. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding, 2024. 4