CounterCurate: Enhancing Physical and Semantic Visio-Linguistic Compositional Reasoning via Counterfactual Examples

Jianrui Zhang[*1] **Mu Cai**[*1] **Tengyang Xie**[1,2] **Yong Jae Lee**[1]

{harrisz,mucai,tx,yongjaelee}@cs.wisc.edu

[1]University of Wisconsin–Madison [2]Microsoft Research

https://countercurate.github.io

Abstract

We propose CounterCurate, a framework to comprehensively improve the visio-linguistic compositional reasoning capability for both contrastive and generative multimodal models. In particular, we identify two critical under-explored problems: the neglect of physically grounded reasoning (counting and position understanding) and the potential of using highly capable text and image generation models for semantic counterfactual fine-tuning. Our work pioneers an approach in addressing these gaps.

We first spotlight the near-chance performance of multimodal models like CLIP and LLaVA in physically grounded compositional reasoning. We then apply simple data augmentation using the grounded image generation model GLIGEN to generate fine-tuning data, resulting in significant performance improvements: +33% and +37% for CLIP and LLaVA, respectively, on our newly curated Flickr30k-Positions benchmark. Moreover, we exploit the capabilities of high-performing text generation and image generation models, specifically GPT-4V and DALLE-3, to curate challenging semantic counterfactuals, thereby further enhancing compositional reasoning capabilities on benchmarks such as SugarCrepe, where CounterCurate **outperforms GPT-4V**.

To facilitate future research, we release our code, dataset, benchmark, and checkpoints at https://countercurate.github.io/.

1 Introduction

Large language models such as ChatGPT (OpenAI, 2023a), GPT4 (OpenAI, 2023b), and LLaMA (Touvron et al., 2023) have demonstrated remarkable knowledge and reasoning abilities. Recently, large multimodal models (LMMs) (Radford et al., 2021; OpenAI, 2023c) further leverage large-scale image-text pairs for improving visio-linguistic understanding capabilities. However, these models exhibit subpar performance in compositional reasoning (Diwan et al., 2022; Hsieh et al., 2023), such as differentiating between semantic distractors like "*white shirts and black pants*" and "*black shirts and white pants*". Current research either concentrates on curating evaluation benchmarks (Diwan et al., 2022; Hsieh et al., 2023; Le et al., 2023) or enhancing compositional reasoning abilities by creating rule-based counterfactual fine-tuning data (Yuksekgonul et al., 2023; Le et al., 2023).

Currently, the majority of research focuses on semantic compositional reasoning, leaving another fundamental problem, physically grounded compositional reasoning, under-explored. This involves tasks such as counting and distinguishing left/right and up/down positions between objects. For example, in Figure 1, GPT-4V (OpenAI, 2023c) made

Figure 1: Representative examples of GPT-4V failure cases. In both questions, GPT-4V correctly identifies all objects in question, but chooses the wrong answer because it fails to distinguish between either left and right (the left question) or up and down (the right question).

[*]Equal Contribution.

the wrong choice even though it identified all objects in question, demonstrating capable semantical yet weak physical compositional reasoning.[1]

With such an observation, it is not surprising to find that both contrastive models like CLIP (Radford et al., 2021) and generative models like LLaVA (Liu et al., 2023b) deliver near random chance performance on our newly curated benchmarks. We hypothesize that modern LMMs are largely oblivious to positional differences. To address this, we generate counterfactual image-text pairs using simple methods, such as horizontal flipping for left/right distinctions, as well as by leveraging a grounded image generation model, GLIGEN (Li et al., 2023), to accurately replace/remove objects of interest for up/down and counting tasks. Our approach shows significant improvements such as 33+% for CLIP and 37+% for LLaVA on our Flickr30k-Positions benchmark.

Existing methods also do not fully utilize the capabilities of powerful generative models when creating semantic counterfactual fine-tuning data. We argue that the key to enhancing compositional reasoning capabilities lies in the careful curation of accurate and sufficiently challenging image and text counterfactuals. For example, augmented negative captions with linguistic rules used by recent fine-tuning approaches (e.g., Yuksekgonul et al., 2023; Le et al., 2023) can unintentionally follow unnatural language distributions, which are easily distinguishable by a text-only model without any image evidence (Lin et al., 2023).

To overcome this, we utilize high-performance text generation model GPT-4V (OpenAI, 2023c) and image generation model DALLE-3 (Betker et al., 2023) to curate reasonably and sufficiently difficult negatives. Our method empirically demonstrates a significant performance boost by fine-tuning CLIP and LLaVA using our data generation pipeline on benchmarks such as SugarCrepe, where we surpass NegCLIP and GPT-4V.

Our contributions are summarized as follows:

- We systematically study physically grounded compositional reasoning, including positional understanding and counting, and highlight the near random performance of multimodal models, including CLIP and LLaVA, on our newly curated benchmarks.
- We significantly improve physical reasoning capabilities by utilizing simple data augmentation techniques and a grounded image generation model, GLIGEN, to generate counterfactual images and captions.
- We employ the most capable image and text generation models to generate semantically counterfactual image/text pairs, further enhancing performance over SOTA methods.

[1]We were able to reproduce our findings of GPT's weakness in multiple fashions: the failure cases themselves were initially discovered via manual testing before the APIs were available; we were able to both use the 2023-07-01 Azure API and the website version of GPT4-V to recreate many of the failure cases as well.

2 Related Work

2.1 Large Multimodal Models

Multimodal models aim at connecting multiple data modalities. For visio-linguistic tasks, multimodal models are constructed either as contrastive (Radford et al., 2021; Jia et al., 2021) or generative (OpenAI, 2023b; Alayrac et al., 2022; Liu et al., 2023b). Recent large multimodal models are trained with large-scale pretraining data and billions of trainable parameters, allowing them to excel at generalization. Contrastive models such as CLIP (Radford et al., 2021) connect images and texts by aligning two modality-specific encoders using contrastive loss. Generative models such as LLaVA (Liu et al., 2023b,a) embed images into a decoder-only language model as prefix tokens and utilize next-token prediction loss to align the two modalities. Though such models demonstrate impressive capabilities across various vision-language tasks such as retrieval and visual question answering, they perform less desirably over compositional reasoning (Diwan et al., 2022; Hsieh et al., 2023).

2.2 Compositional Reasoning

Compositional reasoning (Diwan et al., 2022) in vision and language tasks involves evaluating models' ability to understand and manipulate complex ideas by breaking them down into simpler, constituent components before recombining them in novel ways. A typical example can be distinguishing "blank pants and white shorts" versus "white pants and blank shorts". Winoground (Diwan et al., 2022) is the pioneering benchmark for compositional reasoning, composed of 400 items, each with two images and two corresponding captions. Given an image, multimodal models are asked to find the matching caption from the provided two options, and vice versa. SugarCrepe (Hsieh et al., 2023)

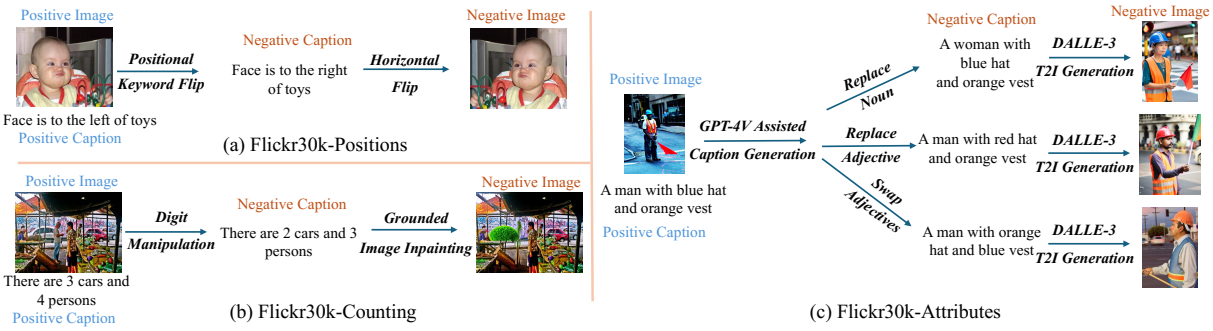


Figure 2: The data curation pipeline of CounterCurate. Given a positive image-caption pair, we first generate the negative captions, based on which we curate the negative images using the most suitable approach. Specifically, (a) for Flickr30k-Positions (left/right), we flip the positional keyword before conducting the horizontal flip for the image; (b) for Flickr30k-Counting, we manipulate the digit before applying grounded image inpainting (Li et al., 2023) as the negative image; (c) for Flickr30k-Attributes, we first leverage GPT-4V (OpenAI, 2023c) to generate reasonable hard negative captions for replacing the noun, replacing the adjective, and swapping the adjectives. Then we leverage DALLE-3 (Betker et al., 2023) to generate coherent images.

and SeeTrue (Yarom et al., 2023) further scale the distractive captions by leveraging language models for automated generation.

NegCLIP (Yuksekgonul et al., 2023) attempts to improve compositional reasoning via fine-tuning CLIP with perturbed captions as the negative text samples in contrastive learning. However, such permuted captions can be easily detected by a text-only transformer (Lin et al., 2023), demonstrating that a more effective approach is needed to further enhance compositional reasoning. DAC (Doveh et al., 2024) leverages dense and aligned captions for better positive captions. SGVL (Herzig et al., 2023) utilizes scene graphs to fine-tune the pre-trained vision-language model. TSVLC (Doveh et al., 2023) leverages language models to generate positive and negative captions. COCO-Counterfactual (Le et al., 2023) further utilizes both negative images and negative captions to fine-tune CLIP via a rule-based data curation pipeline.

Moreover, previous works mainly consider composition reasoning with semantically different concepts rather than physically grounded relationships such as object counting and distinguishing left/right and above/below. Though a recent work (Paiss et al., 2023) improved CLIP's counting capabilities, they neither addressed compositional reasoning nor leveraged negative images. In our paper, we systematically improve both semantic and physical compositional reasoning by leveraging the most capable generative models including GPT-4V (OpenAI, 2023c) and DALLE-3 (Betker et al., 2023) to generate counterfactual image/text pairs.

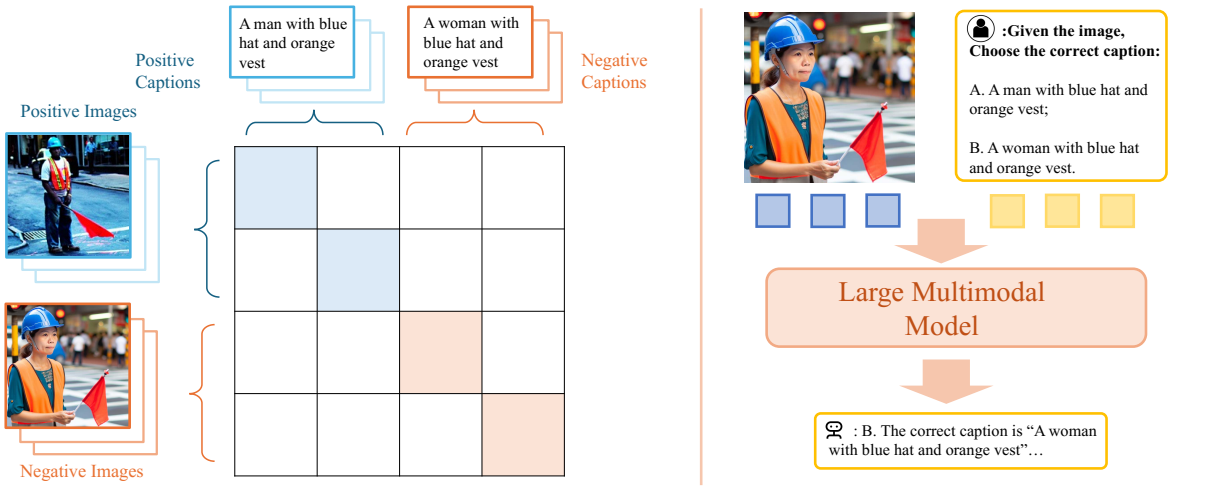
2.3 Counterfactual Reasoning

Counterfactual reasoning (Morgan and Winship, 2015) refers to the process of imagining "what if" scenarios by modifying the input data. In the context of visio-linguistics scenarios, this involves curating negative images and captions by manipulating the original data and observing how it affects the outcome. This helps models understand cause and effect and predict outcomes in situations they've never seen before. The key to counterfactual reasoning is curating meaningful and hard enough examples. COCO-Counterfactual (Le et al., 2023) explores simple linguistic rules to generate negative captions and uses an image editing model Prompt2Prompt (Hertz et al., 2023) to produce negative images (Hertz et al., 2023). Prompt2Prompt is less desirable in understanding complex language instructions, as the generated counterfactuals are less reliable and challenging.

In this paper, we collect the very challenging negative image and text examples by leveraging the most capable language models GPT-4 (OpenAI, 2023b), text-to-image generation model DALLE-3 (Betker et al., 2023), and GLIGEN (Li et al., 2023), significantly improving our model's reasoning. Utilizing these hard negatives, our models can better grasp the nuances of language and vision, enhancing their performance on tasks that require a sophisticated understanding of the world.

3 Approach

In Sec. 3.1, we first explain our selection of the Flickr30k Entities dataset for both benchmarking and fine-tuning. Then in Sec. 3.2 and 3.3, we



(a) Fine-tuning using contrastive learning (CLIP)

(b) Fine-tuning using generative models (LLaVA)

Figure 3: Fine-tuning different types of large multimodal models with CounterCurate. Our pipeline can enhance both contrastive learning models and generative models by augmenting vanilla image-caption pairs with curated negative images and captions. Specifically, our counterfactual image-caption pairs (a) provide auxiliary contrastive loss for models like CLIP, where positive contrastive units in the similarity matrix are colored as blue/red and negative ones are colored as white, and (b) can be naturally integrated into the original next-token prediction loss in text generation models such as LLaVA.

demonstrate the near-chance performance of multimodal models like CLIP and LLaVA on physically grounded compositional reasoning tasks, including positional understanding and counting, and improve their performance by generating negative image-text pairs via data augmentation and grounded image inpainting. Finally, we introduce using capable text and image generation models to improve models' semantic compositional reasoning in Sec. 3.4.

3.1 Necessity of Grounded Image Captions

Previous approaches (Yuksekgonul et al., 2023; Hsieh et al., 2023) adopt global image captioned datasets such as COCO-Captions (Chen et al., 2015) and craft negative captions to either benchmark or improve compositional reasoning. This method, however, lacks regional information important for physically grounded reasoning, such as determining the relative position of two objects. Therefore, we leverage Flickr30k Entities (Plummer et al., 2016), a grounded image captioning dataset to both curate negative image-caption pairs and benchmark models.

3.2 Improving Positional Understanding

Currently, there is no benchmark that directly evaluates the positional understanding of multimodal models. We build such a benchmark, Flickr30k-Positions as in Figure 2 (a), that is composed of positive and negative image-caption pairs describing the same objects in opposite positions.

Generating negative captions. Specifically, we utilize the bounding boxes

$$B = \{B_j = (x_{j1}, y_{j1}, x_{j2}, y_{j2}) \mid \forall j \in I\}$$

in Flickr30k Entities (Plummer et al., 2016) for each object j in Image I, where box B_j's upper-left and lower-right corners are at (x_{j1}, y_{j1}) and (x_{j2}, y_{j2}), respectively. In order to ensure non-ambiguity in relative position descriptions between categories (e.g., "the dog is to the left of the cat"), we only consider images where for each object category, there is only at most one instance of it in the image. For each pair of objects $\{a, b\} \subseteq I$, we identify the positional object pairs via the formula:

$$\begin{cases} x_{a2} \leq x_{b1} \implies a \text{ is to the } \textit{left} \text{ of } b \\ x_{a1} \geq x_{b2} \implies a \text{ is to the } \textit{right} \text{ of } b \\ y_{a2} \leq y_{b1} \implies a \text{ is } \textit{above } b \\ y_{a1} \geq y_{b2} \implies a \text{ is } \textit{below } b \end{cases}$$

With the formula, we generate positive and negative positional captions C and C' such as $C =$ *"a bike is to the left of a woman"* vs. $C' =$ *"a bike is to the right of a women"* and $C =$ *"a table is above a towel"* vs. $C' =$ *"a table is below a towel"*.

We also prompt text-only GPT4 (OpenAI, 2023b) to rewrite the vanilla negative prompt *"a*

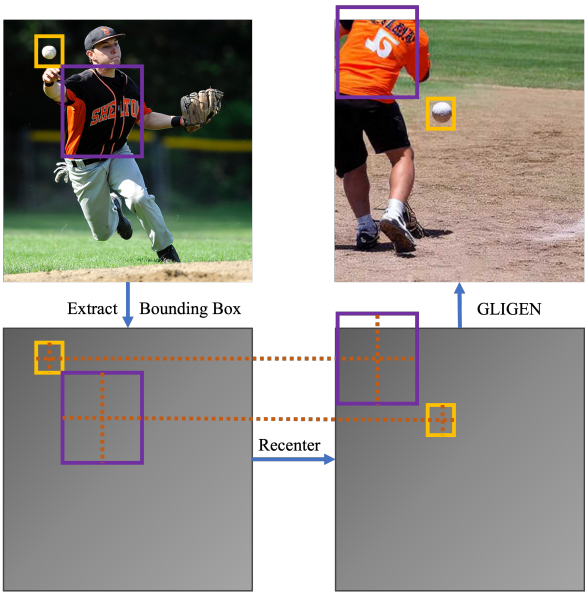


Figure 4: To generate the correct above-below negative image via GLIGEN with the original prompt "*the ball is below the sports outfit*", we recenter the bounding boxes of "*ball*" and "*sports outfit*" and feed them into GLIGEN together with an expanded prompt from GPT4.

is above/below b*"* to better guide image generation models. For example, the caption "a street is above a young child" is unnatural, while the GPT-expanded caption "the street stretches out above a young child, elevated by a bridge" is much better. For detailed prompt engineering, see Appendix B.

Generating negative images. Furthermore, we include negative images in our dataset to curate Flickr30k-Positions. For the left-right pairs, we apply a simple data augmentation method: flip the original images horizontally. For the above-below pairs, however, flipping vertically in most cases would lead to unreasonable images. To address this issue, we utilize GLIGEN (Li et al., 2023), an object-level text-to-image generation model that can ground text prompts with bounding boxes. In particular, for each above-below object pair $(a, b) \in I$, we swap the centers of the bounding boxes B_a, B_b while maintaining their widths and heights, ensuring that the generated images are natural while preserving the objects' sizes and aspect-ratios as demonstrated in Figure 4. The new bounding boxes and prompts are fed to GLIGEN to generate the corresponding negative image I'.

3.3 Improving Object Counting

We introduce Flickr30k-Counting to enhance LMMs' object counting capabilities. We again curate counterfactual image-caption pairs as in Figure 2 (b). We count the number of bounding boxes for each object category in Flickr30k-Entities (Plummer et al., 2016). For two distinct object categories $\{P, Q\} \subseteq I$, we generate a positive caption $C =$*"there are* n_P P*'s and* n_Q Q*'s"*, such as *"there are three cats and four dogs"* where category P is "cat" and category Q is "dog".

Generating negative captions. If all instances of the two selected object categories do not spatially overlap with any other object in the image, we generate the corresponding negative caption by decrementing the one with more objects, Q, and incrementing the one with fewer, P; for example, $C' =$*"there are four cats and three dogs"*. This enforces a hard counterfactual format as the numbers in the positives are reversed in the negatives.

In most cases, however, objects in an image do overlap, and we cannot ascertain that the increment/decrement rule will still generate hard counterfactuals. For example, if a dog overlaps with a cat, and we re-use the rule above to remove one dog and add one cat, we end up with "three cats and three dogs". In this case, we only remove either one of P or one of Q, for example, the dog, and all objects it overlaps with. An example of the resulting negative caption is $C' =$*"there are two cats and three dogs."*

Generating negative images. In the case where objects do not overlap, we simply leverage GLIGEN (Li et al., 2023) to inpaint a new object of the incremented category P at the bounding box of one of the objects from the decremented category Q. Otherwise, we remove the target objects using GLIGEN to replace it with an object from a generic category. We simply use the `plant` category for replacement, as plants can appear in almost any natural setting, let it be outdoor or indoor, and we find that GLIGEN does well in generating it according to each image I's background during inpainting for each corresponding negative image I'. More details can be found in Appendix C.

3.4 Improving Semantic Compositional Reasoning

Existing works (Yuksekgonul et al., 2023; Le et al., 2023) use a rule-based approach to generate negative images/captions for semantic compositional reasoning, resulting in less desirable negatives such as uncommon negative captions e.g., *"A dog with blue fur"* from *"A dog with black fur"*. We introduce Flickr30k-Attributes as in Figure 2 (c), which

utilizes GPT-4V (OpenAI, 2023c) to generate hard negative captions and leverages DALLE-3 (Betker et al., 2023) for generating corresponding high-quality negative images. We extract the first and most detailed caption C for each image I from the original Flickr30k dataset (Young et al., 2014).

Generating negative captions. To generate high-quality negative captions, we leverage GPT-4V and feed both positive images I and captions C to generate negative captions C'. To let GPT-4V know the objects' locations, we overlay bounding boxes B onto the original positive image (Cai et al., 2024) and also feed this annotated image into GPT-4V. We prompt GPT-4V to generate three kinds of captions, including (i) changing a noun, (ii) changing an adjective, and (iii) swapping any of the two phrases, nouns, or adjectives. For example, given the original caption C *"a man wearing a black shirt and blue jeans"*, GPT-4V can generate the following negatives C': (i) *"a man wearing a black* ***jacket*** *and blue jeans"*, (ii) *"a man wearing a black shirt and* ***red*** *jeans"*, and (iii) *"a man wearing a* ***blue*** *shirt and* ***black*** *jeans"*. Detailed prompts to GPT-4V are shown in Appendix A.

Generating negative images. We simply feed the curated captions C' from GPT-4V directly into the currently most capable text-to-image generation model, DALLE-3 (Betker et al., 2023), to generate high-quality negative images I'.

3.5 Fine-tuning Methodology

As demonstrated in Figure 3, CounterCurate can be used to fine-tune both contrastive (Radford et al., 2021) and generative (Liu et al., 2023b) multimodal models. To fine-tune a contrastive model like CLIP, we sample N positive image-text pairs and the corresponding negative image-text pairs to form a batch (C, I, C', I') as shown in Figure 3 (a). We refer to this method as **Grouping**, where the model is forced to distinguish the positive pairs from their corresponding negatives, which is demonstrated to be helpful in our experiments. We also conduct a detailed analysis of the effectiveness of this method in Section 4.5. The training loss is the same as the original, where cross-entropy loss is used to maximize the diagonal elements and minimize all other entries in the similarity matrix.

For generative models like LLaVA (Liu et al., 2023b,a, 2024), we reformat our data into conversations and follow their exact training paradigm to fine-tune, as shown in Figure 3 (b). The training loss is the original next-token prediction loss.

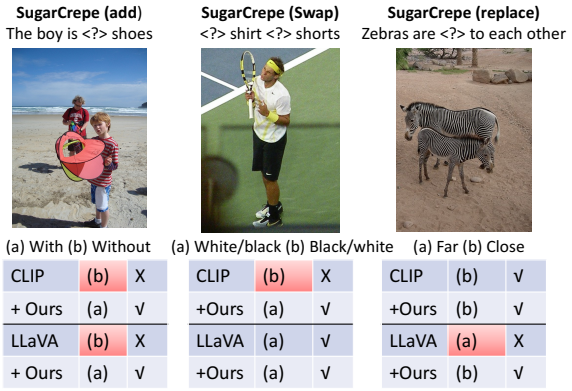


Figure 5: Qualitative examples of models' compositional reasoning capabilities before/after being finetuned via our approach CounterCurate. Wrong answers are marked in red. Our approach enhances both CLIP and LLaVA's reasoning capabilities.

4 Experiments

We utilize the proposed three datasets, Flickr30k-Positions, Flickr30k-Counting, and Flickr30k-Attributes, created via CounterCurate, to fine-tune both contrastive and generative models. Specifically, we select two common multimodal models, ViT-B/32 from (Ilharco et al., 2021) OpenCLIP pre-trained on LAION-2B (Schuhmann et al., 2022) and LLaVA-1.5 (Liu et al., 2023a) as base models. An overview of our results is shown in Figure 5.

4.1 Implementation Details

When generating images for Flickr30k-Attributes, we provide DALLE-3 (Betker et al., 2023) with `hd` quality and `natural` style. We train CLIP (Radford et al., 2021; Ilharco et al., 2021) and LLaVA (Liu et al., 2023a) on 1 and 4 NVIDIA A100 80GB GPUs, respectively. When fine-tuning CLIP, we set Adam optimizer's β_1 and β_2 to be 0.9 and 0.98, and weight decay to be 0.2.

For Flickr30k-Attributes, we train the CLIP model with a learning rate of 1e-5, batch size of 256, and mixed precision for 5 epochs. For LLaVA-1.5, we use a learning rate of 2e-6 and batch size of 16 for 1 epoch. For Flickr30k-Positions, we use a learning rate of 2.56e-5 and batch size of 1024 for CLIP, and train for 50 epochs without grouping and 15 epochs with grouping. For Flickr30k-Counting, we train CLIP with a learning rate of 5e-5. The LLaVA parameters for both Flickr30k-Positions and Flickr30k-Counting are identical, with a learning rate of 2.56e-5 and batch size of 16.

Models	LR	AB	Both
Random	50.00	50.00	50.00
CLIP	50.55	52.80	51.56
+ CounterCurate	**75.88**	**91.52**	**84.90**
LLaVA-1.5	61.84	56.69	59.17
+ CounterCurate	**95.72**	**96.21**	**96.00**

Table 1: Performance of CLIP and LLaVA on Flickr30k-Positions's test dataset. "LR" means left-and-right, and "AB" means above-and-below.

	CLIP	LLaVA-1.5
Vanilla	57.50	44.87
+ CounterCurate	**68.51**	**50.74**

Table 2: Comparison between CLIP and LLaVA on the benchmark created out of PointQA-LookTwice.

4.2 Evaluating Positional Understanding

To the best of our knowledge, we are the first to comprehensively evaluate a multimodal model's ability in understanding the left-and-right and/or above-and-below positioning relations between objects. We conduct a 4:1 train-test split on Flickr30k-Positions and use the test subset as the benchmark. We further separate this dataset into 3 subsets: left-and-right only, above-and-below only, and both. We evaluate our fine-tuned CLIP models by comparing the CLIP similarity scores between the four image-caption pairs, $(C, I), (C, I'), (C', I), (C', I')$ from each group of data in the dataset, where (C, I, C', I') denotes the positive and negative image/caption, respectively. The model receives a score of 0.5 if $s_{(C,I)} > s_{(C',I)}$ and another score of 0.5 if $s_{(C,I')} < s_{(C',I')}$, where $s_{(C,I)}$ is the CLIP cosine-similarity score between caption C and image I. For LLaVA-1.5, we simply query the model to choose between the positive and negative captions when presented with the ground truth image. The results are shown in Table 1.

As we hypothesized, LMMs are indeed largely oblivious to the objects' positioning in the image, which is especially manifested in the vanilla CLIP's performance, which is only marginally better than random guessing. Vanilla LLaVA-1.5 shows only slightly better performance.

After fine-tuning with the training split of Flickr30k-Positions, both models perform significantly better across all subsets. Specifically, for the mixed case, CLIP improves by 33% points, and LLaVA achieves a high accuracy of 96%. These results demonstrate that CounterCurate is highly effective across different kinds of multimodal models.

4.3 Evaluating Object Counting

Here we show the counting capability of our models fine-tuned on Flickr30k-Counting. We select a different dataset, PointQA-LookTwice (Mani et al., 2022), as the evaluation benchmark. Specifically, PointQA-LookTwice was designed to ask models the number n_J of occurrences of an object category J in an image I. We reformat this dataset such that for every J, we generate a positive caption $C_J =$ *"There are n_J J's"* and a negative caption $C'_J =$ *"There are $n_J + 1$ J's"*. For example, given C_J *"there are 3 dogs"*, C'_J is *"there are 4 dogs"*. Similar to what we did in Section 4.2, we mark that a model made a correct prediction if (i) CLIP shows $s_{(C_J,I)} > s_{(C'_J,I)}$ and (ii) LLaVA-1.5 correctly chooses the option for C_J. The results are shown in Table 2.

As CLIP performs slightly better than random guessing, it is surprising that LLaVA-1.5 performs worse than random. Nevertheless, fine-tuning with Flickr30k-Counting improves both models' counting capability. This shows the effectiveness of using GLIGEN-generated negative images in CounterCurate to tackle the problem of counting. We conduct further ablation studies in Appendix F.

4.4 Evaluating Semantic Compositional Reasoning

Similar to the setup in Sec 4.3, we fine-tune both CLIP and LLaVA-1.5 under our curated Flickr30k-Attributes. We use another common benchmark SugarCrepe (Hsieh et al., 2023) to evaluate the model's semantic compositional reasoning capabilities. SugarCrepe has three major categories depicted in Table 3. The evaluation protocol is also the same as Sec 4.3. We choose NegCLIP (Yuksekgonul et al., 2023) and GPT-4V (OpenAI, 2023c) as the representative strong baselines for contrastive and generative models respectively.

We observe significant improvements for both CLIP and LLaVA-1.5, both on average and categorically. Summarized results are shown in Table 3, where Appendix G contains more detailed scores.

For example, CounterCurate fine-tuned CLIP surpasses NegCLIP on average as well as in two main categories. Note that the source of fine-tuning data, Flickr30k-Entities, contains much fewer image-caption pairs than COCO-Captions

Categories	CLIP	NegCLIP	+ CounterCurate	LLaVA-1.5	GPT-4V	+ CounterCurate
Add	80.22	87.28	**89.44** (+4.62)	86.02	91.63	**97.13** (+11.11)
Replace	82.40	85.36	**87.10** (+4.70)	92.38	**93.53**	92.82 (+0.43)
Swap	65.42	**75.31**	72.22 (+6.80)	85.95	88.21	**90.88** (+4.93)
Average	79.54	84.85	**86.15** (+6.61)	89.27	92.19	**94.17** (+4.90)

Table 3: Comparison between performances of CLIP and LLaVA-1.5 on SugarCrepe before and after fine-tuning. We also add NegCLIP (Yuksekgonul et al., 2023) and GPT-4V (OpenAI, 2023c) as strong baselines for contrastive and generative multimodal models. The best performances are bolded, and improvements against CLIP and LLaVA are measured in the parentheses. CounterCurate shows significant performance boost compared to both vanilla CLIP/LLaVA-1.5 model and advanced models such as GPT-4V.

(around 100k image-caption pairs) where NegCLIP is trained. Furthermore, when prompting GPT-4V to generate the negative captions, we intentionally prompt GPT-4V to produce `None` when there is no feasible condition to conduct "swapping". These two factors lead to our data curation pipeline resulting in much fewer negative samples for the "swap" category, which is around 2.5k. We argue that using our pipeline on other datasets that can generate more of the "swap" attribute should lead to larger improvements in performance.

LLaVA-1.5 also performs better in all three categories after fine-tuning. It is also surprising that our fine-tuned model outperforms the SoTA LMM GPT-4V both on average and in two of the categories, most significantly for the "add" category. We observe improvements on other datasets (Diwan et al., 2022) as well in Appendix E. The promising results of LLaVA, such as the score of 94.17% on SugarCrepe surpassing GPT-4V and a 10% improvement on Flickr30k-Position's above-below subset compared to GPT-4V, showcases that CounterCurate effectively improves LLaVA's semantic and physical reasoning capabilities, and we thus identify CounterCurate as a valid candidate towards improving SoTA LMMs such as GPT-4V as well.

CounterCurate shows better performance compared to the prior rule-based methods or less desirable models that generates negatives. These significant improvements show that fine-tuning with accurate and hard negative samples is important, again demonstrating the effectiveness of our data curation and fine-tuning pipeline for both contrastive models and generative models for improving semantic counterfactual understanding.

4.5 In-Depth Analysis

Effectiveness of Negative Images, Negative Captions, and Grouping The core of CounterCurate is to (i) fine-tune models with both negative images and negative captions and (ii) use grouping to help the model better distinguish the positive pairs from the negatives. Here we conduct rigorous ablation studies to demonstrate the necessity of each component. We use the same training parameters and test with the same methods as we did for fine-tuning CLIP and evaluating it on Flickr30k-Positions.

Negative Images	Negative Captions	Group -ing	LR	AB	Both
×	×	×	50.55	52.80	51.56
×	✓	×	55.86	62.25	58.68
✓	×	×	54.35	56.79	54.95
✓	✓	×	69.99	91.24	76.88
✓	✓	✓	**75.88**	**91.52**	**84.90**

Table 4: Ablation study demonstrating the necessity of using (i) negative images, (ii) negative captions, and (iii) grouping strategies. Models are fine-tuned and evaluated on the Flickr30k-Positions dataset.

As shown in Table 4, even though using either the negative caption or the negative image can improve performance, the scores are significantly improved when using both elements to fine-tune. This demonstrates that CounterCurate, which incorporates both negative captions and negative images, is necessary to achieve desirable improvements. Grouping also delivers further improvements compared to not using this strategy.

Correctness of DALLE-3 generated images We randomly sample 300 DALLE-3 images generated from the negative captions in Flickr30k-Attributes for human annotators to check whether the generated image is consistent (matches) with the negative captions. We obtain a consistency score of 84.67%, which demonstrates the high quality of the DALLE-3 generated images.

	MMBench	MMBench_cn	LLaVA_W	POPE	ScienceQA	MM-Vet	VizWiz	MME
LLaVA-1.5	67.7	63.6	70.7	**85.9**	71.6	35.4	53.6	**1531**
+ CounterCurate	**68.6**	**64.1**	**71.1**	85.6	**72.1**	**39.5**	**54.2**	1510

Table 5: Evaluating our fine-tuned LLaVA-1.5 on its original benchmarks.

Scores	CLIP	+ CounterCurate
Image @1	39.44	37.81
Image @5	65.43	64.24
Text @1	56.48	56.96
Text @5	79.74	80.12
Average @1	47.96	47.39
Average @5	72.59	72.18

Table 6: Comparison of image and text retrieval precision scores on the MSCOCO dataset between the original CLIP model and CounterCurate fine-tuned CLIP. The latter is able to maintain overall performance with minor improvements in text retrieval precision.

Performance on zero-shot vision-language tasks We evaluate whether fine-tuning on the counterfactual data hurts the original zero-shot vision-language performance. For CLIP, we compare the image and text retrieval performance on MSCOCO (Lin et al., 2014) of the original CLIP model and CounterCurate fine-tuned model on Flickr30k-Attributes. The results in Table 6 show that on average, the model fine-tuned via CounterCurate maintains performance with marginal difference compared to the original CLIP model. In Appendix I, we also show that fine-tuning with CounterCurate significantly improves CLIP's performance on several common downstream tasks. For LLaVA, we mix Flickr30k-Attributes with LLaVA's original training data before fine-tuning our model. Compared with vanilla LLaVA-1.5 in Table 5 on the same benchmarks from (Liu et al., 2023a), the CounterCurate-finetuned LLaVA overall performs similarly (or even slightly better). From these studies, we see that improving LMMs' semantical compositional reasoning with Flickr30k-Attributes does not hurt downstream performance of both contrastive and generative models.

5 Conclusion

In conclusion, CounterCurate significantly enhances the visio-linguistic compositional reasoning capabilities of multimodal contrastive and generative models. This is achieved by addressing the neglect of physically grounded reasoning and exploiting the potential of using text and image generation models for semantic counterfactual fine-tuning. We believe our contributions can pave the way for further research in compositional reasoning.

Acknowledgement

This work was supported in part by NSF CAREER IIS2150012, and Institute of Information & communications Technology Planning & Evaluation(IITP) grants funded by the Korea government(MSIT) (No. 2022-0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration) and (No. RS2022-00187238, Development of Large Korean Language Model Technology for Efficient Pre-training), and Microsoft Accelerate Foundation Models Research Program.

This research is funded in part by the University of Wisconsin–Madison L&S Honors Program through a Trewartha Senior Thesis Research Grant.

Limitations

We acknowledge that DALLE-3 and GPT4-V are closed-source models, which makes it difficult to systematically analyze their behavior.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS 2022*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. 2023. Improving image generation with better captions.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. 2024. Making large multimodal models understand arbitrary visual prompts. In *CVPR 2024*.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan

Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023. PaLI: A jointly-scaled multilingual language-image model. In *ICLR 2023*.

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

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV 2020*, pages 104–120. Springer.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. 2022. Why is winoground hard? investigating failures in visuolinguistic compositionality. In *EMNLP 2023*.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, et al. 2024. Dense and aligned captions (dac) promote compositional reasoning in vl models. *Advances in Neural Information Processing Systems*, 36.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. 2023. Teaching structured vision & language concepts to vision & language models. In *CVPR 2023*, pages 2657–2668.

Mark Everingham, S M Ali Eslami, Luc Van Gool, Christopher K I Williams, John Winn, and Andrew Zisserman. 2015. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.*, 111(1):98–136.

Andreas Ginger, Philip Lenz, and Raquel Urtasun. 2021. Are we ready for autonomous driving? In *CVPR 2021*.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 12(7):2217–2226.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021. Natural adversarial examples. In *CVPR 2021*. IEEE.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Prompt-to-prompt image editing with cross attention control. *ICLR 2023*.

Roei Herzig, Alon Mendelson, Leonid Karlinsky, Assaf Arbelle, Rogerio Feris, Trevor Darrell, and Amir Globerson. 2023. Incorporating structured representations into pretrained vision & language models using scene graphs. In *EMNLP 2023*.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2023. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *NeurIPS 2023*.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Tiep Le, Vasudev Lal, and Phillip Howard. 2023. Coco-counterfactuals: Automatically constructed counterfactual examples for image-text pairs. In *NeurIPS 2023*.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. Gligen: Open-set grounded text-to-image generation. In *CVPR 2023*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft coco: Common objects in context. In *ECCV 2014*.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. 2023. Visual-gptscore: Visio-linguistic reasoning with multimodal generative pre-training scores. *arXiv preprint arXiv:2306.01879*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. In *CVPR 2024*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *NeurIPS 2023*.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft.

Arjun Mani, Nobline Yoo, Will Hinthorn, and Olga Russakovsky. 2022. Point and ask: Incorporating pointing into visual question answering.

Stephen L Morgan and Christopher Winship. 2015. *Counterfactuals and causal inference*. Cambridge University Press.

OpenAI. 2023a. Chatgpt. https://openai.com/blog/chatgpt/.

OpenAI. 2023b. Gpt-4 technical report.

OpenAI. 2023c. Gpt-4v(ision) system card.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. 2023. Teaching clip to count to ten. In *ICCV 2023*.

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.

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. In *PMLR 2021*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. 2018. Rotation equivariant CNNs for digital pathology.

Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. 2023. What you see is what you read? improving text-image alignment evaluation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions.

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? In *ICLR 2023*.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. 2019. A large-scale study of representation learning with the visual task adaptation benchmark.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *CVPR 2021*, pages 5579–5588.

Appendix

A Prompt guidance to GPT-4V for Flickr30k-Attributes

As a minor detail specific to the dataset itself, we extracted the objectified caption C'_I from the annotated Flickr30k-Entities dataset (Plummer et al., 2016). The objectified captions have square brackets added to the original caption that allow us and the LMMs to refer to each phrase for the object j in the caption with a unique ID $\#_j$.

The exact prompt guidance we provided to GPT-4V (OpenAI, 2023c) is as follows:

You are given an image, the same image but with bounding boxes, its corresponding caption and an enhanced form of the caption. Their format is as follows: Original Caption: A child in a pink dress is helping a baby in a blue dress climb up a set of stairs in an entry way. Enhanced Caption: [/EN#1/people A child] in [/EN#2/clothing a pink dress] helping [/EN#3/people a baby] in [/EN#4/clothing a blue dress] climb up [/EN#5/other a set of stairs] in [/EN#6/scene an entry way]. In the enhanced caption, there is no new data, but that each "entity" is marked by a pair of square brackets. Most entities each correspond to one or more bounding boxes, which will be specified. For example, entity 1 in the sentence is "A child", which is marked by a tag [/EN#1/people ...]. "people" states the type of the entity. If entity is "other", then there are no restrictions applied.

You are tasked to:

Generate a caption that changes the object being discussed in exactly one of the entities. You MUST ensure that the new object is the same type of entity as the original object as specified in the tag. For example: [/EN#1/people A child] => [/EN#1/people An adult] is allowed, but [/EN#1/people A child]

Categories	CLIP	NegCLIP	+ CounterCurate	LLaVA-1.5	GPT-4V	+ CounterCurate
Add Object	77.89	88.78	**90.35** (+3.20)	87.83	91.59	**97.82** (+9.99)
Add Attribute	87.15	82.80	**86.71** (+8.82)	80.64	91.76	**95.09** (+14.45)
Replace Object	93.77	92.68	**95.94** (+2.17)	96.73	96.31	**98.37** (+1.64)
Replace Attribute	82.61	85.91	**87.94** (+5.33)	92.64	**93.53**	**93.53** (+0.89)
Replace Relations	68.92	**76.46**	**76.24** (+7.32)	87.13	**90.26**	85.92 (-1.21)
Swap Object	60.00	**75.51**	68.57 (+8.57)	84.90	83.13	**85.72** (+0.82)
Swap Attribute	67.42	**75.23**	73.57 (+6.15)	86.34	90.09	**92.79** (+6.45)
Average	79.54	84.85	**86.15** (+6.61)	89.27	92.19	**94.17** (+4.90)

Table 7: Detailed version of Table 3: comparison of performance over each sub-category.

=> [/EN#1/people A cat] is not allowed because a cat is not "people";

Generate a caption that changes the qualifier (such as an adjective of a quantifier) that describes the object in exactly one of the entities. For example: [/EN#2/clothing a pink dress] => [/EN#2/clothing a green dress].

Generate, if possible, a caption that reverses two of the entities or their qualifiers such that the original sentence structure is not changed, but produces a negative prompt. For example, given two entities "a green dress" and "a blue blouse", you can either swap the two entities' order or swap the adjectives and produce "a blue dress" and "a green blouse". If you cannot generate one, report None.

All in all, the new description must meet all of these requirements:

1. The change of attribute must be sufficiently different to make the new description inaccurate, but it should also be somewhat related to be challenging to an AI model.

2. Compared to the original description, the new description must differ in only one attribute. All other details must be kept the same.

3. The new description must mimic the sentence structure of the original description.

4. The new description must be fluent, logical, and grammatically correct.

5. Carefully look at the image, and give negative captions that are reasonable given the objects' position, size, and relationship to the overall setting.

6. Pose challenging(difficult enough) negative captions so that a large multimodal text generation model should struggle to distinguish the original caption v.s. negative caption.

Here are some examples whose output format you should follow: Original Caption: A child in a pink dress is helping a baby in a blue dress climb up a set of stairs in an entry way. Enhanced Caption: [/EN#1/people A child] in [/EN#2/clothing a pink blouse] helping [/EN#3/people a baby] in [/EN#4/clothing a blue dress] climb up [/EN#5/other a set of stairs] in [/EN#6/scene an entry way]. Bounding Boxes: #1: purple Your answer: {"noun": {"action": (1, "a child", "an adult"), "caption": "An adult in a green dress is helping a baby in a blue dress climb up a set of stairs in an entry way."]}, "adjective": {"action": (2, "a pink dress", "a green dress"), "caption": "A child in a green dress is helping a baby in a blue dress climb up a set of stairs in an entry way."}, "reverse": {"action": (2, 4), "caption": "A child in a blue blouse is helping a baby in a pink dress climb up a set of stairs in an entry way."}}

B Prompt guidance to GPT-4V for Flickr30k-Positions (above-and-below)

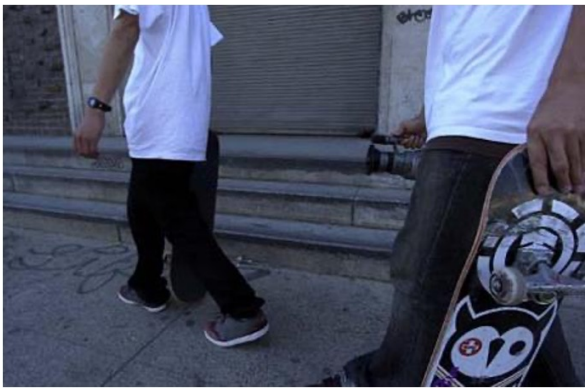
To generate high-quality counterfactual images for the above-and-below subset of Flickr30k-Positions, we use the re-writing technique to generate convincing and detailed captions for GLIGEN (Li et al., 2023). Specifically, we leverage GPT-4V (OpenAI, 2023c) with the following prompt:

I will give you a caption in the format "A is above B." You need to expand the sentence such that the meaning "A is above B" is preserved and your answer is reasonable for a human to understand what you're describing. Do not make the answer too long; one long sentence is enough. For example, if i give you "a man is under a dog", a good answer would be "there is a man resting on the ground, and there is a dog lying above him." One restriction: A and B do not overlap. This means that if I ask you to expand "a hat is below water", you must not assume that the hat is below water. Remember that you MUST include both A and B in your answer, like my example did.

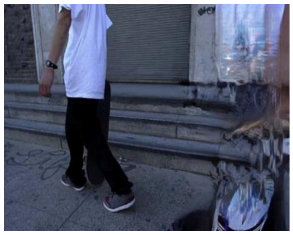
C Object Removal vs Inpainting Plants

We explain here as for why we do not use object removal tools to curate our data for Flickr30k-Counting as we make one example in Figure

6. We used a tool from a public Github Repository https://github.com/treeebooor/object-remove and compared it with GLIGEN (Li et al., 2023)'s inpainting results. We found a significant pattern that when object removal failed, most of GLIGEN's inpainting succeeded.



(a) Original Image



(b) Object Removal

(c) GLIGEN Inpainted Plant

Figure 6: We want to remove/cover the person on the right with their skateboard in image (a) completely. Object removal only achieves the goal partially and makes the image much less natural, while GLIGEN inpainted a plant perfectly into the image as it also preserved the rest of the information in the image.

D Overfitting in Above-Below Subset?

In section 4.2, we observed that both models performed better on the above-and-below subset, with CLIP (Radford et al., 2021; Ilharco et al., 2021) performing significantly better. However, we argue that this is not caused by overfitting because there is an even number of both "A is above B" captions vs. "A is below B" captions, and both captions are not always matched to a generated image by GLIGEN (Li et al., 2023) or the original image from Flickr30k (Young et al., 2014). Since the training and testing datasets are shuffled and separated from the same dataset, this ensures that neither model can find a pattern in, for example, recognizing the generated image apart from the original image and choosing an option on that basis. Another proof is that since LLaVA-1.5 (Liu et al., 2023a) can perform equally well on both subsets after fine-tuning, the two subsets are shown to be at least as well-defined as the other. We also welcome others to use our dataset as a benchmark to test their model's ability to understand object positioning in images.

E Results on the Winoground Dataset

We evaluated our LLaVA-1.5 (Liu et al., 2023a) on Winoground (Diwan et al., 2022), a model specifically made to test a model's limit of visio-linguistic capabilities via human-annoted difficult image-caption pairing questions. Our model, after fine-tuning on Flickr30k-Attributes, showed significant improvements on the text score of Winoground.

Model	Text Score
$\text{CLIP}_{\text{(ViT-B/32)}}$ (Radford et al., 2021)	25.25
$\text{UNITER}_{\text{base}}$ (Chen et al., 2020)	32.25
$\text{UNITER}_{\text{large}}$ (Chen et al., 2020)	38.00
VinVL (Zhang et al., 2021)	37.75
BLIP2 (Zhang et al., 2021)	44.00
PALI (Chen et al., 2023)	46.50
LLaVA-1.5 (Liu et al., 2023a)	65.85
LLaVA-1.5 + CounterCurate	**69.15** (+3.30)

Table 8: Our fine-tuned model of LLaVA-1.5 with Flickr30k-Attributes shows significant improvements on the difficult visio-linguistic reasoning dataset Winoground.

F More Ablations on Flickr30k-Counting

We also test not using any of the generated negative image-caption pairs and not using grouping while fine-tuning CLIP with Flickr30k-Counting. Evaluation results on PointQA-LookTwice are shown in Table 9. Both ablations showed much less improvement, showcasing that using grouping and the negative image-caption pairs is more powerful in improving models' counting abilities.

Model Setting	Accuracy
Vanilla	57.50
+ CounterCurate (No Neg)	60.85
+ CounterCurate (No Group)	65.70
+ CounterCurate	**68.51**

Table 9: More ablations on the CLIP model trained with Flickr30k-Counting. the score without any negatives and the score without grouping.

G More Results on SugarCrepe

Here we show the performance improvements for every sub-category of SugarCrepe in Table 7. Over-

all, CounterCurate shows clear performance gain over the two base models, CLIP (Radford et al., 2021) and LLaVA (Liu et al., 2023b). We also point out that the "swap object" category in SugarCrepe (Hsieh et al., 2023) only has 246 items as compared to other categories' 500+, 1000+ items; this means that the performance on this specific category could show more fluctuations caused by the training process.

H Ratio of Augmentation

We study the level of effectiveness our dataset Flickr30k-Attributes brings upon the models if we choose to reduce the amount of curated data used during fine-tuning. As we observe in Table 10 and especially Table 11, using 100% of our dataset yields the greatest amount of improvements in LLaVA and CLIP's semantical compositional reasoning capabilities. This prompt us to further curate data using the same method in the future that shall bring forth even more competent models.

I More evaluation on CLIP downstream tasks

In Table 12, we compare vanilla CLIP with CounterCurate-finetuned CLIP on Flickr30k-Attributes on some common downstream tasks. The results show that CounterCurate generally improves performance.

Type	LLaVA-1.5	+ CounterCurate	(50%)	(25%)	(10%)	(5%)	(2.5%)
add_att	80.64	**95.09**	94.22	90.90	88.58	88.44	86.27
add_obj	87.83	**97.82**	97.58	95.44	94.67	93.79	90.45
replace_att	92.64	**93.53**	93.27	91.50	92.51	91.62	92.64
replace_obj	96.73	**98.37**	**98.37**	98.24	98.06	97.88	97.28
replace_rel	87.13	85.92	84.92	85.63	86.77	88.12	**88.19**
swap_att	86.34	**92.79**	90.54	89.19	88.44	90.09	89.79
swap_obj	84.90	**85.72**	83.67	81.63	82.04	84.08	**86.12**
add_average	86.02	**97.13**	96.73	94.29	93.13	92.44	89.39
replace_average	92.38	92.83	92.40	92.24	92.79	93.02	93.00
swap_average	85.95	**90.89**	88.69	87.15	86.71	88.47	88.80
Overall	89.27	**94.17**	93.54	92.38	92.18	92.26	91.17

Table 10: Ratio of augmentation ablation study on fine-tuning LLaVA-1.5.

Type	CLIP	+ CounterCurate	(50%)	(25%)	(10%)	(5%)	(2.5%)
add_att	87.14	**90.34**	89.23	88.11	86.85	87.09	87.05
add_obj	77.89	**86.70**	81.21	79.91	78.46	78.03	77.74
replace_att	82.61	**87.94**	84.77	82.99	83.37	82.86	82.61
replace_obj	93.76	**95.94**	94.79	94.37	93.88	93.76	93.76
replace_rel	68.91	**76.24**	72.97	70.19	69.55	69.34	69.13
swap_att	67.41	**73.57**	68.76	66.81	66.66	66.51	67.26
swap_obj	60.00	**68.57**	62.85	62.85	63.26	62.44	60.40
add_average	80.21	**87.62**	83.22	81.97	80.57	80.31	80.08
replace_average	82.39	**87.10**	84.76	83.20	82.83	82.60	82.47
swap_average	65.42	**72.22**	67.17	65.75	65.75	65.42	65.42
Overall	79.54	**85.49**	**85.65**	82.07	80.64	79.94	79.68

Table 11: Ratio of augmentation ablation study on fine-tuning CLIP.

Benchmark	CounterCurate	LAION-2b
ImageNet-a (Hendrycks et al., 2021)	**26.8**	26.25
kitti-closest-vehicle-distance (Ginger et al., 2021)	**30.1**	26.16
eurosat (Helber et al., 2019)	**51.59**	48.13
voc2007_multilabel (Everingham et al., 2015)	**82.49**	79.63
dmlab (Zhai et al., 2019)	**16.81**	15.97
pcam (Veeling et al., 2018)	**60.58**	59.78
fgvc_aircraft (Maji et al., 2013)	**25.26**	24.54

Table 12: Comparing our fine-tuned CLIP with LAION-2b pretrained CLIP on common downstream tasks.