

Enhancing Compositional Reasoning in Vision-Language Models with Synthetic Preference Data

Samarth Mishra

Kate Saenko

Venkatesh Saligrama

Boston University

{samarthm, saenko, srv}@bu.edu

Abstract

Compositionality, or correctly recognizing scenes as compositions of atomic visual concepts, remains difficult for multimodal large language models (MLLMs). Even state of the art MLLMs such as GPT-4o can make mistakes in distinguishing compositions like “dog chasing cat” vs “cat chasing dog”. While on Winoground, a benchmark for measuring such reasoning, MLLMs have made significant progress, they are still far from a human’s performance. We show that compositional reasoning in these models can be improved by elucidating such concepts via data, where a model is trained to prefer the correct caption for an image over a close but incorrect one. We introduce SCRAMBLE: Synthetic Compositional Reasoning Augmentation of MLLMs with Binary preference Learning, an approach for preference tuning open-weight MLLMs on synthetic preference data generated in a fully automated manner from existing image-caption data. SCRAMBLE holistically improves these MLLMs’ compositional reasoning capabilities which we can see through significant improvements across multiple vision language compositionality benchmarks, as well as smaller but significant improvements on general question answering tasks. As a sneak peek, SCRAMBLE tuned Molmo-7B model improves on Winoground from 49.5% to 54.8% (best reported to date), while improving by 1% on more general visual question answering tasks. Code for SCRAMBLE along with tuned models and our synthetic training dataset is available at <https://github.com/samarth4149/SCRAMBLE>.

1. Introduction

Compositionality in vision language models allows them to recognize a scene as a composition of its parts. With this if a model understands atomic concepts, it understands infinite compositions of them without additional representational capacity [24] (see Fig. 1).

A common test of compositionality assumes understand-

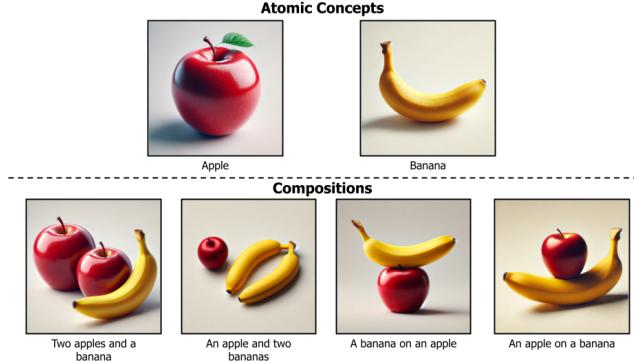


Figure 1. **Compositionality** or recognizing complex scenes as compositions of atomic concepts, is integral to human cognition [24], and therefore desirable in vision language models.

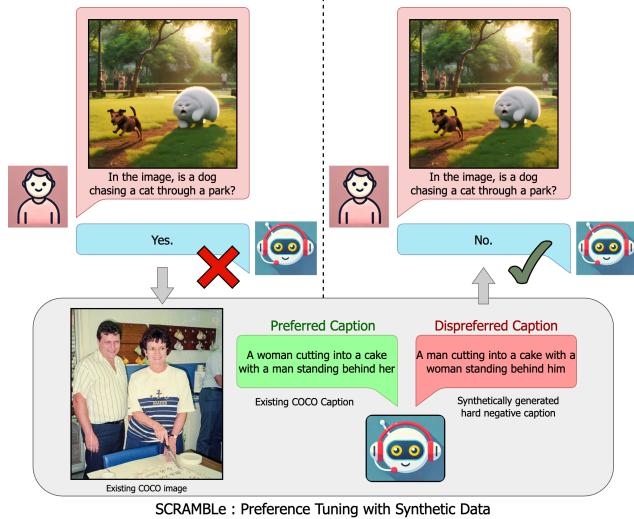


Figure 2. **SCRAMBLE**. Even state of the art multimodal large language models (MLLMs) (like GPT-4o) can falter on tests of compositionality, e.g. determining the correct relation between the dog and the cat in the image. We propose SCRAMBLE, a fully automated approach of collecting synthetic preference data and tuning MLLMs for improving their significantly on compositionality, while also leading to smaller overall improvements as measure by general QA benchmarks.

ing of atomic concepts and assesses if a vision language model can determine how they are composed in a complex scene. Such a test reveals that even state of the art models struggle with compositionality, for *e.g.*, GPT-4o [35] failed at correctly determining the relation between a dog and a cat in the image shown in Fig. 2. To tackle such failures, we introduce SCRAMBLE, a novel approach that automates high-quality synthetic negative caption generation for preference tuning MLLMs. While prior work synthetically generating hard negatives often produce grammatically or logically incorrect captions, a trait models can use to learn shortcuts, we improve quality of hard negatives generated using LLM reasoning and subsequent plausibility and grammar based filtering, ensuring that negatives remain syntactically and semantically valid while still presenting a meaningful challenge for training.

Evaluating Compositionality. Winoground [46] is a seminal benchmark that uncovered such shortcomings in compositionality of large-scale pre-trained vision-language models. A model is given two images and two captions and is asked to match each image to its caption and vice versa. This evaluates compositionality in the manner that the two captions use the same set of words but paint very different pictures, often by changing the relation between objects, but sometimes by changing meanings of words. Besides Winoground, other benchmarks (*e.g.* COLA [42], EqBen [49], ConME [22]) have been used as tests of compositional reasoning focusing on recognition when atomic concepts, their properties or their relations change in complex scenes. Refer to Figs. 5 and 6 for examples from these benchmarks.

Performance of Vision Language Models. Starting with models such as CLIP [39], that were shown to have atrociously low performance (*e.g.* below chance on Winoground [46]), vision-language models have improved significantly by using large language model foundations (as multimodal large language models or MLLMs). For instance, LLaVA [28] has been reported to have a Winoground accuracy (group score) of 30% [27] up from CLIP’s 8%. There is however lots of room for improvement when it comes to compositionality of these models.

Improving Compositionality. Some works have found that better compositionality can be realized by spending more inference time compute with these MLLMs to reason better through the answering process. For instance, Mitra *et al.* [33] and Zhang *et al.* [54] have found different chain of thought prompting methods can help MLLMs improve performance on compositionality benchmarks. Along similar lines, CECE [6] is a method that with the help of an LLM, expands a caption into a larger set of entailments and contradictions, and performs the matching task based on this expanded set. While these methods are effective, a good MLLM should not have to spend time generating such

entailments to answer simple questions about objects comprising an image, their properties and/or relations between them.

SCRAMBLE. With this motivation, we propose to improve MLLMs via preference tuning, to teach them to disprefer failure modes such as complex scene captions that are mostly correct but incorrect in a small/subtle way. While it may be possible to collect many more high quality captions of complex scenes with expensive labeling pipelines to improve MLLM compositionality, we chose this approach since our data curation pipeline is fully automated (hence comparatively inexpensive) using existing image-caption data that MLLMs may have already been trained on (*e.g.* COCO [26]).

Conceptually our approach is similar to NegCLIP [53] where hard negative captions are generated by programmatically swapping nouns/adjectives/verbs/adverb phrases in the original caption for tuning a CLIP model. This approach has also been used for creating evaluation benchmarks [32, 56], but it was Hsieh *et al.* [20] who first pointed that if negative captions are non-grammatical or illogical, models can use shortcuts to beat them. We argue that when such data is used for training, it does not may teach a model only trivial differences between preferred and dispreferred outcomes. Hence, as mentioned above, we focus on high quality hard negative caption generation by using an LLM expert’s reasoning, and filtering to discard low quality (*e.g.* non-grammatical or illogical) captions.

This simple approach substantially improves MLLM performance, not just for compositionality, but holistically. As a sneak peek, on a Molmo-7B [9] model, our SCRAMBLE tuning approach leads to a performance of 54.8% (up from 49.5%), the best reported on Winoground to date. This is achieved along with 1% improvement on general question answering benchmarks such as SEED-Bench [25] and MM-VET [52]. We demonstrate SCRAMBLE’s efficacy using 6 compositionality and general QA benchmarks with 3 recent open-weight MLLMs. We also show SCRAMBLE’s data quality is higher and leads to better performance than prior hard negative caption generation methods.

2. Background and Related Work

Vision Language Models. Large-scale pre-training unlocked many-a-capability in different vision language models [4]. CLIP [39] is one such model with separate image and text encoders trained to match features of images to features of their textual descriptions. This model can then be used for a range of tasks such as zero-shot classification [38, 39, 57], text-to-image retrieval [44], image or text similarity measurement [18], text to image generation [41, 43] among others. CLIP also found another very important application in combination with large language models [8, 47] to create a general purpose image ques-

tion answering chat agent [28, 29, 59]. More capable open weight MLLMs have subsequently followed similar core architectures [2, 9, 15] and closed the gap to proprietary models like GPT-4 [1]. As the more capable class of models, we are interested in their compositionality, which we show can be significantly improved with our approach.

Visio-Linguistic Compositionality. Compositionality is central to human intelligence and allows us to comprehend infinite compositions of finite numbers of concepts [24]. It is hence a desirable in vision language models too. Thrush *et al.* [46] conducted a controlled experiment to test this and found CLIP has worse than chance performance on their dataset, alongside other image and text representation learning approaches [7, 55]. Subsequently Mitra *et al.* [33] and Lin *et al.* [27] proposed ways of evaluating MLLMs on this benchmark (with and without access to model output probability distributions respectively) and found them to be much better than models like CLIP. Diwan *et al.* [10] in their study of Winoground, found some examples may be hard for reasons other than lack of compositional reasoning capability (*e.g.* ambiguity, visual difficulty of identification, etc.). Besides Winoground, we evaluated using other benchmarks like EqBen [49], Cola [42] and ConME [22] for additional compositionality testing scenarios.

Improving Compositionality. After Thrush *et al.*'s [46] study, multiple benchmarks used synthetically generated hard negative captions for evaluating compositionality [32, 53, 56]. Hsieh *et al.* pointed out [20] that if the quality of these hard negatives is poor grammatically or logically, text-only models could exploit this bias to pick the correct caption over the negative one. From these works, Yuksekgonul *et al.* [53] also proposed NegCLIP, *i.e.* tuning CLIP with their generated hard negatives, but this did not lead to much better performance on Winoground (as reported in [17]). Our approach, SCRAMBLE, is similar conceptually to NegCLIP, except with an application to MLLMs and an emphasis on the quality of generated hard negative captions. We compare with a baseline data generation approach based on SugarCREPE [20], a quality-focused evaluation benchmark based on synthetic hard negatives, and find that our generation leads to higher quality hard negatives, and hence to better performing tuned models.

Besides the above, there have been a range of other approaches attempting to improve CLIP's compositionality : Tuning with additional synthetic images and captions [5], with losses to improve equivariance [17], with dense and aligned captions [13], or with structured captions rewritten by LLMs [14]. These approaches can be translated to MLLMs, and may be complementary to our approach. An investigation of them is left for future work.

As mentioned in the introduction, in MLLMs, prior compositionality improvement have been made with additional test time compute to elicit more/better reasoning [6, 33, 54].

SCRAMBLE, in contrast, attempts to improve MLLMs directly by teaching them distinctions with high quality preference data. In this manner, it is complementary to the aforementioned approaches.

Preference Optimization of Language Models. Human preferences over pairs of outcomes are a powerful and relatively easy to collect signal, which has been shown effective in training models for following instructions [36], summarization [45], translation [23] and story-telling [60]. These works trained a separate model to learn a reward function compatible with human preferences, that was then used to fine-tune the language model with reinforcement learning. Rafailov *et al.* [40] proposed a training procedure that bypasses the reward model and fine-tunes a language model directly on preference data. As a stable and computationally efficient training procedure, we use this method for fine-tuning MLLMs with our synthetic preference data.

3. Method

Our approach, SCRAMBLE, introduces a fully automated pipeline for generating high-quality synthetic preference data to improve MLLMs' compositional reasoning. Unlike prior approaches that swap phrases without always checking for coherence (*e.g.* NegCLIP), SCRAMBLE incorporates grammar and plausibility filtering to ensure the generated negatives are both challenging and logically sound.

3.1. Synthetic Data Generation

Baseline : Swap Objects/Attributes. Here we describe a baseline data generation approach similar to SugarCREPE [20] swap. Here, hard negatives are generated by swapping nouns or adjectives in the input caption, when possible. The specific prompt used for swapping nouns is shown in Fig. 3. While an effective method, we found that swaps generated would still often be nonsensical. There were also many cases when two nouns were swapped around phrases like "and", "next to" or "beside" and the resulting negative caption would have the same meaning as the original caption.

Our improved approach with Chain of Thought. In contrast to the baseline, we get the LLM to reason more generally through the negative caption generation process, by providing a chain of thought in the output [50]. The initial prompt to the LLM mentions the same requirements as Fig. 3, with the distinction that we now ask the LLM to rearrange words as it sees fit. To help with the reasoning process, we provide the LLM expert with in-context examples. One such example (out of the 5 we provide) is in Fig. 4. We source these examples also automatically by choosing 5 caption pairs (no images) from Winoground and asking Claude 3.5 Sonnet [3] how it would reason through if it was given the task of transforming one caption to its counterpart under the 3 requirements in Fig. 3.

Given an input sentence describing a scene, your task is to first locate two swappable noun phrases in the sentence, and then swap them to make a new sentence. The new sentence must meet the following three requirements:

1. The new sentence must be describing a different scene from the input sentence.
2. The new sentence must be fluent and grammatically correct.
3. The new sentence must make logical sense.

To complete the task, you should:

1. Answer the question of whether generating such a new sentence is possible using Yes or No.
2. Output the swappable noun phrases.
3. Swap the selected noun phrases to generate a new sentence.

Input: A woman cutting into a cake with a man standing behind her.

Is it possible to swap noun phrases in the input sentence to generate a new sentence that is different from the input sentence and makes logical sense? Yes.

Swappable noun phrases: a woman, a man

Output: A man cutting into a cake with a woman standing behind him.

Figure 3. **Baseline Hard-negative Generation : Swap Objects.** Example prompt to swap noun phrases and generated output (green) from the Llama-3.1 expert.

3.2. Filtering

We note the negative caption generation method we described is noisy and susceptible to making errors. Hence, we use a post filtering step to keep only a subset of high quality negative captions. We use the Vera plausibility model [30] and the grammar model from TextAttack [34] in an adversarial refinement process (similar to [20]) for this step. This process is aimed at debiasing the generated dataset based on the score differences between positive and negative captions. It ensures there are an equal number of examples with a certain score difference as there are with the negative of that score difference, *i.e.*, the distribution of score differences is symmetric about zero. This debiasing prevents a model trained on this data from learning shortcuts using the plausibility or grammatical correctness of the captions (and disregarding images). For the sake of brevity, the adversarial refinement procedure is described in full detail in Appendix B.

We also note that while adversarial refinement debiases the dataset, it also acts as a filter of low quality captions in terms of grammar and plausibility. Most positive captions have fairly high plausibility and grammar scores, given the images are captioned by humans. Hence, if negative captions with low scores were to be retained, to keep the dataset

Input Caption: a pink bird with a white beak
Reasoning:

1. Identify the key elements:
 - Color of the bird: pink
 - Color of the beak: white
 - Structure: "[color] bird with a [color] beak"
2. Recognize that the negative caption should describe a different visual image using the same words
3. Observe that the colors are the main distinguishing features
4. Swap the colors while maintaining the structure:
 - "pink" describing the bird becomes "white"
 - "white" describing the beak becomes "pink"
5. Keep the rest of the structure intact:
 - "a [color] bird with a [color] beak"
6. Apply the swapped colors to the structure:
 - "a white bird with a pink beak"

Final Output Caption : a white bird with a pink beak

Figure 4. **A Reasoning Chain of Thought** for transforming a caption into a negative caption. We provide 5 such examples in-context to the Llama-3.1 LLM expert to help it reason through the process of generating a new negative caption.

unbiased, an equal number of examples would be needed with the negative score difference, *i.e.*, examples where the negative caption is a lot more plausible and grammatically correct compared to the positive caption. Such examples are extremely rare and in some cases impossible because the upper bound on the scores is 1.

3.3. Training

Once we have a preference dataset (\mathcal{D}) of images, their captions and hard negative captions, we fine-tune a low-rank adapter [21] on a target MLLM with the direct preference optimization [40] objective:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right]. \quad (1)$$

where x is a prompt, in our case, an image and the prompt "caption : ", y_w is the preferred outcome *i.e.* the image caption, and y_l is the dispreferred hard negative caption. We use the MLLM before finetuning as the reference model π_{ref} . π_θ is the model with the adapter (θ being the trainable adapter parameters). β is a parameter controlling the deviation from the reference model and σ the sigmoid function. Note that $\pi(y | x)$ is the probability of generating the caption y given the prompt x from the model π .

4. Experiments

4.1. Evaluation.

We evaluated the performance of the tuned MLLMs on four compositionality benchmarks (Winoground [46], EqBen [49], Cola [42] and ConMe [22]) and two other image question answering benchmarks (SEED-Bench [25] and MM-Vet [52]), used as control to ensure the tuned MLLMs do not lose performance on more general tasks.

Compositionality Benchmarks. An example from ConMe consists of 1 image based on which the MLLM is asked a question with 2 options denoted with “A” or “B”, and it is expected to generate the letter corresponding to the correct option (see examples in Fig. 6).

The remaining three benchmarks involve a matching task with 2 images and 2 captions in each example (see examples in Fig. 5). Winoground examples could involve object swaps, relation swaps or both. While the first reorders elements such as noun phrases referring to objects in the image, relation swaps reorder elements such as verbs, adjectives, prepositions, and/or adverbs. Swaps of both objects and relations could lead to meanings of words changing across the captions. EqBen is a dataset consisting of both real and synthetic images, with typically a small aspect changed between two images. Cola has examples from natural images with different relations between two objects. An example is deemed correctly answered if for each image, the model prefers its caption over the other and vice versa for each caption (this corresponds to the group score accuracy defined in [46]).

Image-Caption Affinity metric (VQAScore). In the compositionality benchmarks with 2 images and captions each, a model can be scored if it can provide a preference/affinity metric between an image and a caption. For this, we use VQAScore [27], where a model’s score for an image-caption pair is the probability, $\pi_\theta(y | x)$, of generating the output $y = \text{“Yes”}$ when asked the question $x = \text{“Does this image show } \langle \text{caption} \rangle \text{?”}$. Note that this requires the output probability distribution over a model’s vocabulary tokens which is not typically available for closed-source models.

Control Benchmarks. SEED-Bench [25] is a large evaluation set of multiple choice questions over both images and videos, of which we only use the image subset containing 14.2k questions. It contains questions testing a variety of capabilities such as scene understanding, instance identity/attribute identification, counting and visual reasoning among others. MM-Vet [52] is a dataset of 218 questions where the model has to generate a descriptive answer. The generated answer is compared with a ground truth answer by a GPT-4 judge and scored between 0 and 1. The questions test abilities like recognition, optical character recognition, general knowledge, spatial reasoning and math.

4.2. Implementation Details

We ran all experiments on single 48G Nvidia Ampere gpus. For the LLM expert generating synthetic captions, we used a Llama-3.1-70B model with inference run at 4-bit (nf4) quantization¹. We tuned and evaluated LLaVA-v1.5-13B [28]², Molmo-7B-D-0924 [9] and Llama-3.2-11B-Vision-Instruct [15] (where 13B, 7B and 11B refer to the approx. number of parameters of the models). We added a low-rank adapter (LoRA [21]) to each model and trained at half (bf16) precision for 2 epochs with an effective batch size of 8 (using gradient accumulation). For full training details please refer to Appendix D.

4.3. Results

SCRAMBLE improves performance of different MLLMs over compositionality and control benchmarks. We experimented with SCRAMBLE tuning three recent open-weight MLLMs and the results are reported in Tab. 1. For Winoground, EQBen and Cola, we report only the group score accuracy and for ConMe and SEED-Bench, we report average accuracy overall. For MM-Vet performance we report the average and 95% confidence interval over 5 evaluations with a GPT-4 judge [52].

All models improve in performance across most compositionality and general QA control benchmarks, showing that SCRAMBLE tuning is effective across MLLMs and can holistically improve their performance while focusing on compositionality. We note that while Molmo seems to become worse on EQBen, on qualitative evaluation on examples where Molmo and SCRAMBLE-Molmo disagree, we find that the former actually makes errors in recognition that the latter does not make (*e.g.* in Fig. 7). On examples where SCRAMBLE-Molmo performed worse, we found no difference in recognition capabilities on interacting via chat (examples in Appendix C). We also note Llama-3.2 model’s performance on SEED-Bench is quite low primarily because it often tended to explain solutions and did not follow the format instructions of answering precisely for the benchmark. While this highlights a format misalignment rather than a compositionality failure, it leaves open the question of whether better prompting strategies could improve performance without additional fine-tuning. As a second note, while tuning Llama-3.2, we found that it overfits when trained with the full dataset of 57.8k synthetic examples, but does improve when tuned on a smaller set (~10k examples). More discussion/results on this in Appendix D.2.

Comparing Caption Generation Procedures. We show

¹we found little performance difference between the quantized and non-quantized models

²We experimented with LLaVA-v1.6 with both vicuna-13B and mstral-7B versions but found them to have poorer compositionality than LLaVA-v1.5



Figure 5. **Examples from compositionality benchmarks.** Winoground has examples in natural images swapping objects, relations or both. EqBen has both natural and synthetic images in examples where typically a small aspect changes between two images. Cola has examples from natural images with different relations between two objects.

Model Name	Compositionality Benchmarks				Control Benchmarks	
	Winoground	EqBen	COLA	ConMe	SEED-Bench	MM-Vet
LLaVA-1.5-13B	36.5	36.4	49.5	62.3	68.23	36.2 ± 0.3
+ SCRAMBLE	39.3	39.3	55.7	64.5	68.19	38.6 \pm 0.1
MoLMo-7B	49.5	62.9	57.1	72.2	74.04	59.3 ± 0.2
+ SCRAMBLE	54.8	59.3	60.5	74.6	74.61	60.9 \pm 0.4
Llama-3.2-11B	31.5	43.6	37.1	71.3	13.79	57.0 ± 0.1
+ SCRAMBLE*	35.3	44.3	40.0	74.6	42.74	60.3 \pm 0.1

Table 1. **SCRAMBLE improves performance of different open-weight MLLMs across compositionality and control benchmarks.** We note that LLaMA-3.2 shows significantly weaker performance on SEED-Bench, largely due to formatting issues rather than an inherent lack of capability. Refer to Sec. 4.3 for full discussion.



Figure 6. **Examples from ConMe.** ConMe is another compositionality test benchmark that poses a two choice question to an MLLM, asking it to output A or B for its choice.

our chain of thought based hard negative generation leads to higher quality synthetic data than the baseline swap objects/attributes approach. Note that each method first involves prompting a Llama-3.1-70B LLM expert to get an initial set of captions, followed by adversarial filtering (Sec. 3.2). In the first stage, the LLM is given an option to

abstain from generation if it considers the task to be impossible.

The number of examples we end up with from each generation method at the end of these two stages is shown in Tab. 3. While looking at the two generation methods of swapping objects and attributes, we notice that the LLM expert attempts to swap objects much more often than attributes. This however leads to a high number of low quality captions that get filtered out. We found that we could improve the initial generation quality a bit by using in-context examples (as done in [20] and as we do for swapping attributes) but this typically resulted in many more abstentions, even when a swap of noun phrases was possible. We find that using reasoning with chain of thought, initial generations have a higher quality and a lot fewer of them are discarded in the filtering stage.

Some examples of generated captions are shown in Tab. 2. In the first three, we see that attributes or objects are swapped as expected by the baseline method. Our chain of thought approach generated the same outcome. In the fourth, using the baseline method of swap objects, the LLM expert generates a non-sensical outcome. In an example like

Positive Caption	Generated Negative Caption	
	(Baseline) Swap Obj/Att	SCRAMBLE : Chain of Thought
A red fire hydrant, a yellow balloon, and some rocks. A woman cutting into a cake with a man standing behind her A young child and two adults go skiing. A young zebra is sniffing the ground in a dusty area. Two giraffes standing in a tree-filled area.	(Att) A yellow fire hydrant, a red balloon, and some rocks (Obj) A man cutting into a cake with a woman standing behind him (Obj) A young adult and two children go skiing. (Obj) A young area is sniffing the ground in a dusty zebra. (Att) NA	A yellow fire hydrant, a red balloon, and some rocks A man cutting into a cake with a woman standing behind him. Two children and a young adult go skiing An old zebra is sniffing the ground in a dusty area. Two giraffes standing outside a tree-filled area.

Table 2. **Qualitative examples** of negative captions generated by baseline vs our approach (red highlights abstentions/poor generations). Swap works well in certain cases when swaps are possible (first three). Chain of thought can handle cases outside of this with higher quality hard negatives.

Stage	(Baseline) Swap Obj	Swap Att	Chain of Thought
Start		118287	
Initial Generation	110855	21813	107679
Post-Filtering	15546	7904	57786
Final		23450	57786

Table 3. **Number of examples** initially generated and those retained after adversarial refinement for each negative caption generation method.

the last, with the swap attributes directive, the LLM expert chose to abstain. In both these cases, our chain of thought based method generates a good quality hard negative that is meaningful and grammatical while changing few words.

In Tab. 4, we compare a SCRAMBLE tuned LLaVA-1.5 model (tuned with DPO with synthetic data generated using our chain of thought method) with to a LLaVA-1.5 model and one that we tuned using the baseline data generated with the swap directives. While the baseline data generation approach leads to some improvements on compositionality benchmarks, we find that it degrades performance on long question answering (as measured by MM-Vet performance) immensely. We also found that for long answers it can on occasion degenerate into repeating a single character/phrase in the output. On the other hand, we found that with SCRAMBLE tuning improves performance across the board, including on long-question answering on MM-Vet.

Winoground Performance. In Tab. 5 we compare SCRAMBLE against prior reported compositionality performance on Winoground. In the top section are the performances of a human and a random chance baseline. The next section contains performances of different image and text representation models. CLIP [39] as reported by Thrush *et al.* [46] performs poorer than random chance, and attempts at improvement by training on compositional data and/or with auxiliary losses [5, 17] could not improve beyond this. METER [12] and FIBER [11] models break past random chance improving with training on additional data with equivariant constraints [49]. In the section below, we report performance of GPT-4V [1] with different prompting meth-

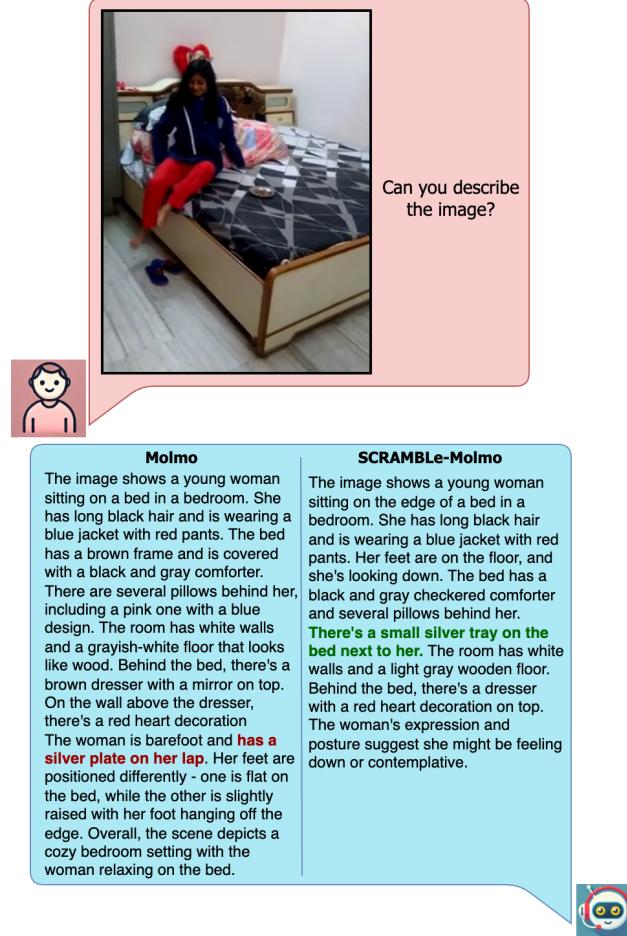


Figure 7. **Conversing with SCRAMBLE-Molmo (EQBen example).** SCRAMBLE can fix mistakes Molmo makes in identifying the location of the silver plate.

ods [33, 54]³. Since GPT-4 is a proprietary model, its output probability distributions have not been made available and thus computing VQAScore [27] is not possible. In the last section, we use this evaluation with different methods on

³We note that the evaluation methods of these two prompting approaches are different since the former [33] restricts itself to models with a single image input while the latter [54] uses multiple images in context

Model Name	Tuning Data	Compositionality Benchmarks				Control Benchmarks	
		Winoground	EqBen	COLA	ConMe	SEED-Bench	MM-Vet
LLaVA-1.5-13B	-	36.5	36.4	49.5	62.3	68.23	36.2 ± 0.3
Baseline (w tuning)	Swap Obj/Att	38.8	36.4	52.9	64.4	68.49	30.7 ± 0.4
SCRAMBLE (Ours)	Chain of Thought	39.3	39.3	55.7	64.5	68.19	38.6 ± 0.1

Table 4. **SCRAMBLE vs baseline caption generation.** We find the chain of thought method to lead to better quality hard negatives and hence better final performance.

Model Name	Winoground		
	Text	Image	Group
Mturk Human [§]	89.5	88.5	85.5
Random Chance [§]	25.0	25.0	16.7
Image-Text Encoder Models			
CLIP (ViT-B/32) [§] [39]	30.8	10.5	8.0
SynCLIP [5]	30.0	11.5	9.5
CLIP-SGVL [17]	32.0	14.0	9.8
METER [†] [12]	39.3	15.8	12.0
METER+EQSIM [49]	45.0	22.8	18.8
FIBER [†] [11]	46.3	25.8	22.2
FIBER+EQSIM [49]	51.5	31.5	27.5
Proprietary Models + Prompting Methods			
GPT4V [‡] [1]	60.3	45.3	33.5
GPT4V-CCoT [33]	64.0	54.5	43.3
GPT4V+CoCoT [54]	58.5	49.5	44.5
Open MLLM Methods + VQAScore Evaluation			
LLaVA-1.5-13B [28]	51.5	50.5	36.5
CECE (LLaVA-1.6) [6]	52.0	61.3	42.8
CECE (LLaVA-1.5+LLaVA-1.6) [6]	55.0	61.3	47.5
Molmo-7B [9]	62.8	61.8	49.5
SCRAMBLE-LLaVA-1.5-13B (Ours)	53.3	52.8	39.3
SCRAMBLE-Molmo-7B (Ours)	66.8	66.3	54.8

Table 5. **Winoground Performance.** Comparison of SCRAMBLE tuned models with prior work. SCRAMBLE-Molmo improves significantly on previous best. ([†]reported in [49], [‡]reported in [33], [§]reported in [46])

open models⁴ and find that SCRAMBLE preference tuning with Molmo-7B-D-0924 leads to much higher performance than previously reported on Winoground.

Conversing with SCRAMBLE-Molmo. We show two examples (more in Appendix C) of chatting with Molmo and SCRAMBLE-Molmo. The first (Fig. 7) is an example from EQBen where we find SCRAMBLE-Molmo correctly determines the location of the plate relative to the woman, in contrast to Molmo’s mistake. In the second example

⁴On evaluating LLaVA-1.5-13B with VQAScore, we find that predicting the probability of “Yes” as opposed to “Yes⟨\s⟩” leads to better performance (where ⟨\s⟩ is the end of sentence token). We believe this is also more principled since it measures the probability of the model starting its answer with “Yes” as opposed to simply giving the one word answer “Yes.”



Figure 8. **Conversing with SCRAMBLE-Molmo (Winoground example).** SCRAMBLE fixes Molmo’s mistake identifying what the man holds.

(Fig. 8), we see that SCRAMBLE-Molmo correctly identifies the man is holding a mitt and not a hose, which is what Molmo perceives.

5. Conclusion

We introduced SCRAMBLE, a procedure of improving visio-linguistic compositionality of multimodal large language models (MLLMs), by tuning them on synthetic preference data. Our synthetic data uses existing image caption data, along with synthetically generated hard negative captions. The generation procedure is fully automated, using an open-weight LLM expert (Llama-3.1-70B) prompted to transform positive captions into hard negative captions by providing its working in a chain of thought. Tuning state of the art open weight MLLMs with this synthetic preference data significantly improves their compositional reasoning performance while also leading to improvements in general question answering (albeit to a smaller extent).

Acknowledgements

The authors Venkatesh Saligrama and Samarth Mishra were supported by the Army Research Office Grant W911NF2110246, AFRLGrant FA8650-22-C1039, the National Science Foundation grants CPS-2317079, CCF2007350 and CCF-1955981.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3, 7, 8
- [2] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Devendra Chaplot, Jessica Chudnovsky, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. 3
- [3] AI Anthropic. Introducing the next generation of claudie, 2024. 3
- [4] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 2
- [5] Paola Cascante-Bonilla, Khaled Shehada, James Seale Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gul Varol, Aude Oliva, Vicente Ordóñez, Rogerio Feris, et al. Going beyond nouns with vision & language models using synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20155–20165, 2023. 3, 7, 8
- [6] Paola Cascante-Bonilla, Yu Hou, Yang Trista Cao, Hal Daumé III, and Rachel Rudinger. Natural language inference improves compositionality in vision-language models. In *International Conference on Learning Representations (ICLR)*, 2025. 2, 3, 8
- [7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 3
- [8] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6, 2023. 2
- [9] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 2, 3, 5, 8
- [10] Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. Why is winoground hard? investigating failures in visuolinguistic compositionality. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, Dublin, Ireland, 2022. Association for Computational Linguistics. 3
- [11] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *Advances in neural information processing systems*, 35:32942–32956, 2022. 7, 8
- [12] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuhang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022. 7, 8
- [13] Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, et al. Dense and aligned captions (dac) promote compositional reasoning in vl models. *Advances in Neural Information Processing Systems*, 36:76137–76150, 2023. 3
- [14] Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2668, 2023. 3
- [15] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3, 5
- [16] Hugging Face. Hugging face hub: A platform for sharing machine learning models, datasets, and spaces. <https://huggingface.co/docs/hub/en/index>, 2024. Accessed: YYYY-MM-DD. 14
- [17] Roei Herzig, Alon Mendelson, Leonid Karlinsky, Assaf Arbelle, Rogerio Feris, Trevor Darrell, and Amir Globerson. Incorporating structured representations into pretrained vision & language models using scene graphs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, 2023. Association for Computational Linguistics. 3, 7, 8
- [18] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. 2
- [19] Matthew Honnibal and Ines Montani. spacy: Industrial-strength natural language processing in python, 2015. 12
- [20] 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, 2024. 2, 3, 4, 6
- [21] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen.

- Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 4, 5
- [22] Irene Huang, Wei Lin, Muhammad Jehanzeb Mirza, Jacob A. Hansen, Sivan Doveh, Victor Ion Butoi, Roei Herzig, Assaf Arbelle, Hilde Kuehne, Trevor Darrell, Chuang Gan, Aude Oliva, Rogerio Feris, and Leonid Karlinsky. Conme: Rethinking evaluation of compositional reasoning for modern vlms. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*, 2024. 2, 3, 5
- [23] Julia Kreutzer, Joshua Uyheng, and Stefan Riezler. Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, 2018. Association for Computational Linguistics. 3
- [24] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017. 1, 3
- [25] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 2, 5
- [26] 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. 2
- [27] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024. 2, 3, 5, 7
- [28] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2, 3, 5, 8
- [29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3
- [30] Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. Vera: A General-Purpose Plausibility Estimation Model for Commonsense Statements. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, 2023. Association for Computational Linguistics. 4, 12, 14
- [31] Shihong Liu, Samuel Yu, Zhiqiu Lin, Deepak Pathak, and Deva Ramanan. Language models as black-box optimizers for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12687–12697, 2024. 12
- [32] 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, 3
- [33] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431, 2024. 2, 3, 7, 8
- [34] John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 119–139, Online, 2020. Association for Computational Linguistics. 4, 12, 14
- [35] OpenAI. Hello gpt-4o, 2024. Accessed: 2024-11-06. 2
- [36] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 3
- [37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 14
- [38] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701, 2023. 2
- [39] 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, 7, 8
- [40] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 4, 14
- [41] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 2
- [42] Arjjit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan Plummer, Ranjay Krishna, and Kate Saenko. Cola: A benchmark for compositional text-to-image retrieval. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 5
- [43] 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

- [44] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314, 2023. 2
- [45] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020. 3
- [46] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 2, 3, 5, 7, 8
- [47] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2
- [48] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. <https://github.com/huggingface/trl>, 2020. 14
- [49] Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11998–12008, 2023. 2, 3, 5, 7, 8
- [50] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 3
- [51] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chau-mond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, 2020. Association for Computational Linguistics. 14
- [52] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 2, 5
- [53] 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, 3
- [54] Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. *arXiv preprint arXiv:2401.02582*, 2024. 2, 3, 7, 8
- [55] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021. 3
- [56] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. An explainable toolbox for evaluating pre-trained vision-language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 30–37, Abu Dhabi, UAE, 2022. Association for Computational Linguistics. 2, 3
- [57] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2
- [58] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning, 2024. 14
- [59] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3
- [60] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. 3

Appendices

A. Alternative Data Generation Approach : Feedback Loop

We experimented with a third method of synthetic data generation but found the data quality to be poorer than that generated using chain of thought. For completeness, we report the method and experiments here. In this approach, we provide feedback to the LLM expert in context in an attempt to get it to refine the negative caption that it generated. We provide feedback along 4 different dimensions :

- **Plausibility** : We use the Vera model [30] to score how plausible a generated caption is, from 0 to 1. An illogical/nonsensical caption would have a low score, while a caption that is plausible would score higher.
- **Grammar** : We use the grammar model from TextAttack [34] to score how grammatical a generated caption is, from 0 to 1. Lower scores indicate poorer grammar.
- **Distinction** : This is a binary response which is 1/Yes if the generated caption is visually distinct from the original caption, and 0/No otherwise. We use a different Llama-3.1 expert that determines this given the original caption and the generated caption.
- **New and Missing Words** : We lemmatize all the words in the original and the new caption using spaCy [19] and in the feedback mention which extra words the LLM has used and which words from the original caption are missing in the generated caption.

The start of an example conversation is in Fig. 9. Since there is no one measure of the quality of a hard negative caption, we attempt to get the LLM to optimize each of the scores the best it can. In this setup, the LLM thus acts as a black-box optimizer [31].

The initial prompt to the LLM expert mentions the same requirements as in the Swap-Objects prompt (Appendix E.1), with the distinction of asking the expert to rearrange words in the input caption as it sees fit. After the LLM generates a caption, we run the feedback models and provide the scores to the expert in the same context. We repeat this process for 5 iterations and pick the caption which is judged distinct from the input and has the highest average score over plausibility, grammar and word similarity (proportional to the Jaccard similarity between the sets of words in the two captions : a score of 1 indicating the same set of words were used in the two captions). A full example conversation is in Appendix E.4.

Results. The results of tuning a base LLaVA-1.5-13B model on data generated with this approach are in Tab. 6 (denoted as Baseline-II). We find that this method does well on some compositionality benchmarks (COLA and ConMe) but is not consistently better than the base LLaVA model especially on the control benchmarks. Qualitative exam-

```
Given an input caption describing a scene,  
your task is to rearrange words in it to make  
a new caption. The new caption must meet the  
following three requirements:  
1. It must describe a scene with visual  
differences to the input caption.  
2. It must be fluent and grammatically  
correct.  
3. It must make logical sense.  
You can choose to abstain and output 'NA'  
if it is not possible to generate a negative  
caption for the given input.  
To help with your task, I will rate your  
output based on grammar (0-1), plausibility  
(0-1), and whether there are visual  
differences between the original caption and  
your output (Yes/No).  
Here is the input caption: A white horse  
pulling a cart down a street.  
New Caption: A cart is being pulled down the  
street by a white horse.  
FEEDBACK:  
Your grammar score is 0.99.  
Your plausibility score is 0.56.  
Is the output caption visually different from  
the original caption? : No  
Your output caption has extra words  
(lemmatized): 'the', 'by', 'be'.  
Can you please try again?  
:
```

Figure 9. **Feedback Loop.** Example prompt to generate a negative caption (generated caption in green) and feedback generated using auxiliary models (in blue). The Llama-3.1 expert is provided the feedback and prompted to try again for 5 iterations.

ples of generated hard negatives from this approach and from the baseline swap objects/attributes approach along with SCRAMBLE’s chain of thought approach are in Tab. 7. We found that the feedback loop method could handle some more complex cases where a logical swap is not possible, but still the quality of generated captions is poorer than SCRAMBLE’s chain of thought approach.

B. Adversarial Refinement

In Sec. 3.2 we described the adversarial refinement procedure to filter out examples for debiasing the preference tuning dataset using grammar and plausibility scores. The goal of this is that only based on plausibility or grammar scores of the captions (while disregarding the image) a model should not be able to correctly guess the positive caption over the negative (at any more than 50% accuracy). We find that this debiasing is also effective for the preference tuning dataset, to avoid any model fitting to these biases. Algorithm 1 shows the adversarial refinement procedure.

		Compositionality Benchmarks				Control Benchmarks	
Model Name	Tuning Data	Winoground	EqBen	COLA	ConMe	SEED-Bench	MM-Vet
LLaVA-1.5-13B	-	36.5	36.4	49.5	62.3	68.23	36.2 ± 0.3
Baseline (w tuning)	Swap Obj/Att	38.8	36.4	52.9	64.4	68.49	30.7 ± 0.4
Baseline-II (w tuning)	Feedback Loop	37.5	33.6	57.1	65.4	67.78	36.3 ± 0.2
SCRAMBLE (Ours)	Chain of Thought	39.3	39.3	55.7	64.5	68.19	38.6 \pm 0.1

Table 6. **SCRAMBLE vs other caption generation methods.** Adding to Tab 4 from the main paper, we report the results of tuning the base LLaVA-1.5-13B with synthetic data from the synthetic data generated using a feedback loop. We find that this method does well on some compositionality benchmarks(COLA and ConMe) but is not consistently better than the base LLaVA model especially on the control benchmarks.

Positive Caption	Baseline : Swap Obj/Att	Baseline-II : Feedback Loop	SCRAMBLE : Chain of Thought
A white horse pulling a cart down a street.	(Obj) A white cart pulling a horse down a street. (Att) NA	A cart is being pushed by a white horse up a street. Close-up of bins of food that exclude broccoli and include bread.	A white horse pushing a cart down a street. Wide shot of bins of food that include fruits and desserts
A truck is pulling a horse trailer at a festival.	(Obj) A horse is pulling a truck trailer at a festival.	A horse is pulling a trailer instead of a truck at a festival.	A truck is being loaded with a horse trailer at a festival.
Two women and a man posing for a photo on the dance floor.	(Obj) Two men and a woman posing for a photo on the dance floor.	Two women and a man posing for a photo off the dance floor.	One woman and two men posing for a photo on the dance floor.
A yellow and blue fire hydrant that has fallen into the street with "caution" tape around it.	(Att) A blue and yellow fire hydrant that has fallen into the street with "caution" tape around it.	A yellow and blue fire hydrant stands upright that has not fallen into the street with "caution" tape around it	A yellow and blue fire hydrant near the sidewalk with 'caution' tape blocking off part of the street.
A man with a white dress shirt black tie and a beard.	(Att) A man with a black dress shirt white tie and a beard.	A man with a black dress shirt and a white tie.	A man with a black dress shirt white tie and no beard.
Three men in military suits are sitting on a bench.	(Obj) Three benches are sitting on military suits.	Three men in military suits are walking past a bench.	Three men in military suits are standing on a bench.
A cake decorated to look like a female mouse.	(Obj) A mouse decorated to look like a female cake.	A cake decorated to look like a male mouse.	A cake decorated to look like a male mouse.
Black and white photo of a man on the sidewalk pulling his luggage.	(Att) White and black photo of a man on the sidewalk pulling his luggage.	A man is pushing his luggage on the sidewalk in a black and white photo.	Color photo of a man on the sidewalk pushing his luggage.
A yellow and white bed in a small room.	(Obj) A small bed in a yellow and white room.	A white bed in a large room with yellow walls.	A yellow and white bed in a large room.

Table 7. **Additional Qualitative Examples** of negative captions generated by different methods.

LLaVA-1.5-13B Perf	Winoground	EqBen	COLA
w adv ref (9.8k e.g.)	38.5	34.3	44.8
w/o adv ref (16.7k e.g.)	32.0	31.4	43.3

Table 8. **Ablating Adversarial Refinement.** When our preference data has not been filtered using adversarial refinement, the performance of LLaVA-1.5-13B drops significantly on the compositionality benchmarks.

In Tab. 8 we show the performance of LLaVA-1.5-13B with and without adversarial refinement. We carry out this experiment by training the LLaVA-1.5-13B model on a smaller set of 16.7k examples from the COCO train set. After running adversarial refinement, we are left with 9.8k examples. Comparing performances of the two models, we

see that tuning with the unfiltered data, causes performance on the compositionality benchmarks to drop significantly, indicating that adversarial refinement is crucial for retaining high quality examples for compositionality learning.

C. Conversing with SCRAMBLE-Molmo : More Examples

More examples of conversive with Molmo and SCRAMBLE-Molmo are in Figs. 10 to 14. Please check the corresponding captions for more details.

D. Implementation Details

All experiments in the paper were conducted on single Nvidia Ampere GPUs with a minimum 48G of VRAM

Algorithm 1 Adversarial Refinement

Require: Grammar model M_G and plausibility model M_P ; Number of grids K ; A set of candidates $\mathcal{D} = \{I_i, T_i^P, T_i^n\}_{i \in [N]}$, where I_i , T_i^P , and T_i^n are i -th image, positive caption, and negative caption.

Ensure: A subset $\bar{\mathcal{D}} \subset \mathcal{D}$

- 1: Calculate the model score gap for each candidate $g_i^{(1)} = M_G(T_i^P) - M_G(T_i^n)$ and $g_i^{(2)} = M_P(T_i^P) - M_P(T_i^n)$.
- 2: Split the 2D space $[-1, 1] \times [-1, 1]$ to $K \times K$ equal-size grids.
- 3: Place each candidate to a grid based on the score gaps $g_i^{(1)}$ and $g_i^{(2)}$.
- 4: Initialize $\bar{\mathcal{D}} = \{\}$
- 5: **for** each pair of grid (G_j, G_j^*) symmetric about the original point $(0, 0)$ **do**
- 6: **if** $|G_j| > |G_j^*|$ **then**
- 7: Sample $|G_j^*|$ candidates from G_j and put them to $\bar{\mathcal{D}}$.
- 8: Put candidates in G_j^* to $\bar{\mathcal{D}}$.
- 9: **else**
- 10: Sample $|G_j|$ candidates from G_j^* and put them to $\bar{\mathcal{D}}$.
- 11: Put candidates in G_j to $\bar{\mathcal{D}}$.

(A100/A6000/A40/L40S/L40/RTX6000ada). We used the PyTorch framework [37] and our code for training MLLMs is based on Huggingface Transformers [51], TRL [48] and POVID [58]. We will upload our tuned models along with our synthetic data to Huggingface hub [16] along with the public release of our work.

D.1. Synthetic Data Generation.

As the LLM expert for synthetic caption generation, we used the Meta-Llama-3.1-70B-Instruct model. We ran inference at 4-bit quantization(nf4), with top-p sampling ($p=0.9$) and a temperature of 0.2. As our auxiliary grammar model we used textattack [34] and as the plausibility model we used Vera [30]. These models were used both for filtering as well as feedback generation.

D.2. Training

We used the direct preference optimization(DPO) [40] objective for preference tuning, as described in Sec. 3.3. As prescribed by Rafailov *et al.*, we used a β value of 0.1. We trained each model with the AdamW optimizer, a base learning rate of $1e-5$ and a cosine learning rate schedule with linear warmup for 3% of the steps.

LLaVA. For tuning the LLaVA-1.5-13B model, we trained for 2 epochs at a batch size of 8 (with no gradient accumulation). The rank of the low rank adapter (LoRA) was set to 32, with the α parameter set to 64. The base learning rate was $1e-5$ and for the projector connecting the visual encoder to the language model we used a learning rate $2e-5$. The first stage of training took between 1.5-3 days to run (depending on the gpu used). For the second stage, we used a batch size of 1 and with 8 steps of gradient accumulation

(for the effective batch size of 8; batch size was reduced to 1 because of the large memory footprint of some of the long LLaVA instruction tuning examples). We trained for 2 epochs with the same learning rates as stage 1. To prevent overfitting we used a label smoothing value of 0.1 in the DPO loss in this stage. This stage took 5-10 hrs to run.

Molmo. We trained the Molmo-7B-D-0924 model for 2 epochs at a batch size of 2 and 4 gradient accumulation steps (for the same effective batch size of 8). The rank of LoRA was set to 16, with the α parameter set to 32. This took 1-1.5 days to train.

Llama-3.2. We trained the Llama-3.2-11B-Vision-Instruct model for 1 epoch at a batch size of 4 with 2 gradient accumulation steps (for the same effective batch size of 8). The rank of LoRA was set to 32, with the α parameter set to 64. In Sec. 4.3 we mentioned that we found this model to overfit to the full set of 57.8k synthetic examples. We hence trained this on a smaller set with 9.8k examples. This took around 12 hrs to train on an Nvidia A40 GPU.

In Tab. 9 we show the performance of the LLama-3 model on being trained with the full set of 58k synthetic examples. While the VQAScore evaluation on compositionality benchmarks still improves over the original Llama-3.2-11B-Vision-Instruct model, a benchmark like MM-Vet revealed some degenerate behaviors. Specifically, in this long answer generation task, the model trained on the full synthetic set often fell into loops of repeating a single phrase or a character, leading to drastic reduction in performance. We also note that the Llama-3.2 model performs poorly on SEED-Bench because it does not follow the format of the benchmark (responding to a multiple choice question with the letter corresponding to the correct answer) even when prompted to do so. This behavior improves a bit with our tuning, while even in this case, tuning with the smaller set is better.

E. Synthetic Data Generating Conversations

Here we show examples of the specific prompts/conversations with the LLM expert for the different methods of generation. In each of the conversations, the output of the LLM is colored in green. For the feedback loop, the output of the auxiliary feedback models is colored in blue.

E.1. Baseline : Swap Objects.

Given an input sentence describing a scene, your task is to first locate two swappable noun phrases in the sentence, and then swap them to make a new sentence. The new sentence must meet the following three requirements:

1. The new sentence must be describing a different scene from the input sentence.

Model	Winoground	EqBen	COLA	ConME	SEED-Bench	MM-Vet
Llama-3.2-11B	31.5	43.6	37.1	71.3	13.79	57.0 ± 0.1
+SCRAMBLE (57.8k eg)	34.3	43.6	33.8	70.1	27.79	35.0 ± 0.4
+SCRAMBLE (9.8k eg)	35.3	44.3	40.0	74.6	42.74	60.3 ± 0.1

Table 9. Llama-3.2-11B-Vision-Instruct on being tuned with the full set of 57k synthetic examples overfits to training data and leads to poorer performance across benchmarks.

2. The new sentence must be fluent and grammatically correct.
 3. The new sentence must make logical sense.
 To complete the task, you should:
 1. Answer the question of whether generating such a new sentence is possible using Yes or No.
 2. Output the swappable noun phrases.
 3. Swap the selected noun phrases to generate a new sentence.
 Input: A woman cutting into a cake with a man standing behind her.
 Is it possible to swap noun phrases in the input sentence to generate a new sentence that is different from the input sentence and makes logical sense? Yes.
 Swappable noun phrases: a woman, a man
 Output: A man cutting into a cake with a woman standing behind him.

different from the input sentence and makes logical sense? Yes
 Swappable attributes: pink, blue
 Output: A girl in a blue shirt holding a pink umbrella.
 Input: A car and a truck are going through the intersection.
 Is it possible to swap attributes in the input sentence to generate a new sentence that is different from the input sentence and makes logical sense? No
 Swappable attributes: NA
 Output: NA
 Input: A cold drink on a hot day.
 Is it possible to swap attributes in the input sentence to generate a new sentence that is different from the input sentence and makes logical sense? Yes
 Swappable attributes: cold, hot
 Output: A hot drink on a cold day.
 Input: Four yellow airplanes flying side by side at an air show.
 Is it possible to swap attributes in the input sentence to generate a new sentence that is different from the input sentence and makes logical sense? No
 Swappable attributes: NA
 Output: NA
 Input: The dress on the left is long and the dress on the right is short.
 Is it possible to swap attributes in the input sentence to generate a new sentence that is different from the input sentence and makes logical sense? Yes
 Swappable attributes: long, short
 Output: The dress on the left is short and the dress on the right is long.
 Input: A big grey elephant standing in the jungle.
 Is it possible to swap attributes in the input sentence to generate a new sentence that is different from the input sentence and makes logical sense? No
 Swappable attributes: NA
 Output: NA
 Input: A large black dog is lying on a white comforter.
 Is it possible to swap attributes in the input

E.2. Baseline : Swap Attributes.

Given an input sentence describing a scene, your task is to first locate two swappable adjectives in the sentence describing different objects, and then swap them to make a new sentence.
 The new sentence must meet the following three requirements:
 1. The new sentence must be describing a different scene from the input sentence.
 2. The new sentence must be fluent and grammatically correct.
 3. The new sentence must make logical sense.
 To complete the task, you should:
 1. Answer the question of whether generating such a new sentence is possible using Yes or No.
 2. Output the swappable adjectives.
 3. Swap them to make a new sentence.
 Here are some examples:
 Input: A girl in a pink shirt holding a blue umbrella.
 Is it possible to swap attributes in the input sentence to generate a new sentence that is

sentence to generate a new sentence that is different from the input sentence and makes logical sense? Yes

Swappable attributes: black, white

Output: A large white dog is lying on a black comforter.

Input: A red fire hydrant, a yellow balloon, and some rocks.

Is it possible to swap attributes in the input sentence to generate a new sentence that is different from the input sentence and makes logical sense? Yes

Swappable attributes: red, yellow

Output: A yellow fire hydrant, a red balloon, and some rocks.

This transformation maintains the same words and grammatical structure but inverts the color assignments, creating a visually distinct image that serves as an effective negative caption.

Final Output Caption: a white bird with a pink beak

Input Caption: a bottle is in water
Reasoning:

1. Identify the key elements:
 - Object 1: bottle
 - Object 2: water
 - Relationship: "is in"
2. Recognize that the negative caption should describe a different visual image using the same words
3. Observe that the relationship between the bottle and water is the distinguishing feature
4. Swap the positions of "bottle" and "water" in the sentence:
 - "Bottle" moves from the subject to the object position
 - "Water" moves from the object to the subject position
5. Adjust the articles as needed:
 - "a bottle" remains "a bottle" (no change needed)
 - "water" remains "water" (no article needed for this uncountable noun)
6. Keep the relationship "is in" intact
7. Reconstruct the sentence with the swapped elements:
 - "water is in a bottle"

This transformation maintains the same core words but inverts the spatial relationship between the bottle and water. It creates a contrasting scenario that effectively serves as a negative caption, describing a completely different visual image while using the same words.

Final Output Caption: water is in a bottle

Input Caption: Manning a ship.

Reasoning:

E.3. SCRAMBLE : Chain of Thought

Given an input caption describing a scene, your task is to rearrange words in it to make a new caption.

The new caption must meet the following three requirements:

1. It must describe a scene with visual differences compared to the scene described by the input caption.
2. It must be fluent and grammatically correct.
3. It must make logical sense.

Note that you can choose to abstain and output 'NA' if it is not possible to generate a negative caption for the given input.

Here are some examples:

Input Caption: a pink bird with a white beak
Reasoning:

1. Identify the key elements:
 - Color of the bird: pink
 - Color of the beak: white
 - Structure: "[color] bird with a [color] beak"
2. Recognize that the negative caption should describe a different visual image using the same words
3. Observe that the colors are the main distinguishing features
4. Swap the colors while maintaining the structure:
 - "pink" describing the bird becomes "white"
 - "white" describing the beak becomes "pink"
5. Keep the rest of the structure intact:
 - "a [color] bird with a [color] beak"
6. Apply the swapped colors to the structure:
 - "a white bird with a pink beak"

Input Caption: Manning a ship.

Reasoning:

1. Identify the key words: "manning" and "ship"
2. Recognize that "manning" is a verb form derived from "man"
3. Swap the roles of "man" and "ship":
 - "Man" becomes the object
 - "Ship" becomes the basis for the verb
4. Transform "ship" into a verb:
 - "Ship" becomes "shipping"
5. Change "man" to its object form:
 - "Man" becomes "a man"
6. Rearrange the words to form a grammatically correct sentence:
 - "Shipping a man"

This transformation maintains the same core words but completely changes the meaning and visual imagery, creating a contrasting scenario that serves as an effective negative caption.
Final Output Caption: Shipping a man.

Input Caption: there are three bananas and two apples

Reasoning:

1. Identify the key elements:

- Object 1: bananas
- Object 2: apples
- Quantity of bananas: three
- Quantity of apples: two
- Structure: "there are [number] [fruit] and [number] [fruit]"

2. Recognize that the negative caption should describe a different visual image using the same words

3. Observe that the quantities are the main distinguishing features

4. Swap the quantities while maintaining the fruits they describe:

- "three" describing bananas becomes "two"
- "two" describing apples becomes "three"

5. Keep the rest of the structure intact:

- "there are" at the beginning
- "and" connecting the two fruit phrases

6. Apply the swapped quantities to the structure:

- "there are two bananas and three apples"

This transformation maintains the same words and grammatical structure but inverts the quantities assigned to each fruit. It creates a visually distinct image that serves as an effective negative caption, describing a different arrangement of fruits while using the exact same words as the positive caption.

Final Output Caption: there are two bananas and three apples

Input Caption: There are more skiers than snowboarders.

Reasoning:

1. Identify the key elements:

- Group 1: skiers
- Group 2: snowboarders
- Comparative structure: "more [group] than [group]"

- Opening phrase: "There are"

2. Recognize that the negative caption should describe a different visual image using the same words

3. Observe that the comparative relationship between skiers and snowboarders is the distinguishing feature

4. Swap the positions of "skiers" and "snowboarders" in the sentence:

- "skiers" moves from the subject of comparison to the object

- "snowboarders" moves from the object of comparison to the subject

5. Keep the comparative structure "more ... than" intact

6. Maintain the opening phrase "There are"

7. Reconstruct the sentence with the swapped elements:

- "There are more snowboarders than skiers"

This transformation maintains the same words and grammatical structure but inverts the comparative relationship between skiers and snowboarders. It creates a contrasting scenario that effectively serves as a negative caption, describing a completely different visual image (with snowboarders outnumbering skiers instead of vice versa) while using the exact same words as the positive caption.

Final Output Caption: There are more snowboarders than skiers.

Input Caption: A white horse pulling a cart down a street.

Reasoning:

1. Identify the key elements:

- Object 1: horse
- Color of the horse: white
- Object 2: cart
- Action: pulling
- Location: down a street

2. Recognize that the negative caption should describe a different visual image using the same words

3. Observe that the action and the relationship between the horse and cart are the main distinguishing features

4. Swap the roles of the horse and cart:

- The horse is no longer pulling the cart
- Instead, the cart could be pulling the horse (although this is less common, we can use a different verb to make it more plausible)

5. Change the verb to make the new scenario more believable:

- Instead of "pulling", use "being pulled by"

6. Keep the rest of the structure intact:

- "A [color] [object] [verb] a [object] down a [location]"

7. Apply the changes to the structure:

- "A white cart being pulled by a horse down a street" is not the best option, as it changes the color and type of the object being pulled. A better option would be to change the direction or the action.

Alternative solution:

- Change the direction:

- "A white horse pulling a cart up a street"

- Change the action:
 - "A white horse pushing a cart down a street"
- Final Output Caption: A white horse pushing a cart down a street.

E.4. Baseline-II : Feedback Loop

Given an input caption describing a scene, your task is to rearrange words in it to make a new caption. The new caption must meet the following three requirements:

1. It must describe a scene with visual differences compared to the scene described by the input caption.
2. It must be fluent and grammatically correct.
3. It must make logical sense.

Note that you can choose to abstain and output 'NA' if it is not possible to generate a negative caption for the given input.

To help with your task, I will rate your output based on grammar (0-1), plausibility (0-1), and whether there are visual differences between the original caption and your output (Yes/No).

In your output, please follow the format

Final Output Caption: <caption>.

Input: A white horse pulling a cart down a street.

Final Output Caption: A cart is being pulled down the street by a white horse.

FEEDBACK:

Your grammar score is 0.99.

Your plausibility score is 0.56.

Is the output caption visually different from the original caption? : No

Your output caption has extra words (lemmatized): 'the', 'by', 'be'.

Can you please try again?

Final Output Caption: A cart is pulled by a horse down the street.

FEEDBACK:

Your grammar score degraded to 0.99.

Your plausibility score improved to 0.88.

Is the output caption visually different from the original caption? : No

Your output caption has extra words (lemmatized): 'the', 'by', 'be'.

Your output caption has missing words (lemmatized): 'white'.

Can you please try again?

Final Output Caption: A horse is not pulling a cart down a street.

FEEDBACK:

Your grammar score degraded to 0.98.

Your plausibility score degraded to 0.24.

Is the output caption visually different from the original caption? : Yes

Your output caption has extra words (lemmatized): 'not', 'be'.

Your output caption has missing words (lemmatized): 'white'.

Can you please try again?

Final Output Caption: A white horse is standing next to a cart on a street.

FEEDBACK:

Your grammar score improved to 0.99.

Your plausibility score improved to 0.34.

Is the output caption visually different from the original caption? : Yes

Your output caption has extra words (lemmatized): 'next', 'stand', 'to', 'on', 'be'.

Your output caption has missing words (lemmatized): 'down', 'pull'.

Can you please try again?

Final Output Caption: A cart is being pushed by a white horse up a street.

FEEDBACK:

Your grammar score improved to 0.99.

Your plausibility score improved to 0.45.

Is the output caption visually different from the original caption? : Yes

Your output caption has extra words (lemmatized): 'up', 'push', 'by', 'be'.

Your output caption has missing words (lemmatized): 'down', 'pull'.



Can you describe this image?

Molmo

This image captures a delightful scene of three children enjoying a swim in a clear blue pool. They're all wearing colorful swimming goggles and are positioned on a large black and blue inner tube.

On the left is a girl with long black hair in a ponytail, wearing blue goggles and a black swimsuit. She's smiling and looking at the boy next to her.

The boy in the middle has short black hair and is wearing green goggles with blue lenses. He's dressed in orange swim trunks with white stripes and is looking back at the girl.

To the right is a younger girl with short black hair, wearing white goggles. She's smiling and looking directly at the camera.

The water is incredibly clear, allowing you to see the bottom. The scene is bright and sunny, creating a cheerful atmosphere perfect for a fun day at the pool. The children appear to be having a great time together, creating a heartwarming and joyful image.



Can you describe the image?

Molmo

The image shows a kitchen scene focused on cooking. There's a black stovetop with a silver burner, and on it is a black frying pan. The pan contains a mixture of ingredients including red and white items, possibly vegetables or meat, along with some green herbs. A hand is visible in the upper right corner, pouring water from a clear bottle with a white cap into the pan. The bottle has a label that says "5 cups water" in white text. The countertop is light brown, and in the background, you can see a light blue wall. In the bottom right corner, there's text that says "TARLA DALAL" in blue and green letters. Overall, it appears to be a cooking demonstration or recipe being filmed, with the chef adding water to the pan of ingredients.

SCRAMBLE-Molmo

The image shows a close-up view of someone cooking on a stovetop. There's a black frying pan on one of the burners with what appears to be a mixture of chopped vegetables and meat. Someone's hand is visible, pouring water into the pan. The stovetop is silver with black grates, sitting on a light brown countertop. In the bottom right corner, there's text that says "TARLA DALAL" in blue and green. Overall, it looks like a cooking demonstration or tutorial.

Figure 10. **Conversing with SCRAMBLE-Molmo (Winoground example).** In this example (which SCRAMBLE-Molmo gets right but Molmo does not), only SCRAMBLE-Molmo correctly identifies that only two of the kids are on the inflatable and the third one is not.

Figure 11. **Conversing with SCRAMBLE-Molmo (EQBen example).** The image is from an EQBen example that Molmo gets right but SCRAMBLE-Molmo does not. However, in conversation, SCRAMBLE-Molmo correctly identifies all elements of the scene and answers the question no worse than Molmo.



Can you describe the image?



Molmo

The image shows a scene on a grassy field with three main objects:

1. A white teddy bear lying on its back in the center. The bear has a red shirt and a white tail.
2. A black and silver bucket with a black handle to the left of the bear.
3. A pair of red-handled scissors with silver blades to the right of the bucket.

The grass appears to be a mix of green and brown, suggesting it's not very well-maintained. The scene is viewed from above, giving a clear top-down perspective of the objects. The teddy bear's position, with its back on the ground and red shirt visible, creates an unusual and somewhat melancholic atmosphere. The presence of the bucket and scissors adds an element of mystery to the scene, as if something out of place has occurred.

SCRAMBLE-Molmo

The image shows an outdoor scene with three objects on grass. There's a small white stuffed elephant lying on its side, a black and silver bucket with a handle, and a pair of red-handled scissors. The grass appears to be a mix of green and brown. The stuffed elephant has a red object in its mouth. The bucket has a logo that looks like a shield with a lion. The scissors are open. It's an interesting arrangement of everyday items in a natural setting.



Is there something odd about the looks of the picture itself? Does it look naturally captured?

SCRAMBLE-Molmo

Yes, there are several aspects of this picture that give it an unusual and somewhat unnatural appearance. The most striking oddity is the quality of the image. While the objects in the scene - the bucket, scissors, and stuffed animal - appear realistic, the overall image has a very artificial look to it.

The grass, in particular, seems to be the main culprit. Instead of looking natural, it has an unnaturally perfect texture. The blades appear to be too uniformly aligned, lacking the organic variations you'd expect in real grass. This uniformity gives the grass an almost plastic-like quality, which is at odds with how grass typically grows in nature.

Additionally, the lighting and shadows in the image don't seem to comport with what you'd expect in a natural outdoor setting. They appear slightly off, contributing to the overall artificial feel of the picture.

These elements combine to create an image that, while containing realistic objects, has an overall look that's more reminiscent of a highly rendered 3D scene or a heavily processed photograph rather than a naturally captured image. The contrast between the seemingly real objects and the unnatural quality of the background creates a somewhat disconcerting effect.



Does the image look odd?



Molmo

Yes, the image does look odd. The scene appears to be from a video game, which gives it an artificial quality. The objects in the image - a teddy bear, a bucket, and scissors - seem out of place and arranged in a way that's more typical of game environments than real life. The grassy background and the slightly off proportions of the objects contribute to the surreal, game-like atmosphere. This unusual composition creates a sense of unease or strangeness, which is often a deliberate effect in certain types of video games or artistic representations.

SCRAMBLE-Molmo

Yes, the image does look odd. The scene presents an unusual combination of objects that don't typically go together in this way. We have a small, white stuffed elephant lying on the ground next to a bucket and a pair of scissors. This odd trio of items in a grassy setting creates an strange and somewhat surreal atmosphere. The contrast between the innocent, childlike stuffed animal and the more utilitarian objects like the bucket and scissors adds to the oddity of the scene. It's the kind of image that might make you question what story or scenario it's trying to convey.





Figure 12. **Conversing with SCRAMBLE-Molmo (EQBen example).** This is another EQBen example that Molmo gets right but SCRAMBLE-Molmo does not. SCRAMBLE-Molmo does correctly identify the elements of the image. As a further question, when asked if there is something odd about the image, Molmo directly answers that the image looks artificially generated and game like, which SCRAMBLE-Molmo picks on the fact that the objects in the image do not typically appear together and it starts to read into the meaning of this. When probed further about the looks of the image however, it does mention that the image looks 3D rendered.



Molmo
The image features a striking neon cactus sculpture with a distinctive design. The cactus has three arms, with the central arm extending straight up and the other two curving outward. **What makes this cactus unique is that the central arm is shaped like a middle finger, adding a rebellious or provocative element to the design.**

The neon cactus is illuminated in a vibrant green color, creating a stark contrast against the purple background. It's mounted on a circular base, which appears to be purple as well, matching the overall color scheme of the scene.

The sculpture is set in what looks like a desert landscape, with purple sand and rock formations surrounding it. This creates an otherworldly, surreal atmosphere that complements the unconventional design of the cactus.

A black cord extends from the base of the sculpture, likely for power or support. The entire scene is bathed in a purple glow, enhancing the futuristic and artistic feel of the image.

This neon cactus sculpture combines elements of nature, art, and pop culture in a bold, eye-catching display that's sure to make a statement in any space.

SCRAMBLE-Molmo
This image features a striking neon cactus with an interesting twist. **The cactus has the typical arms and shape of a neon sign, but instead of the usual friendly "hello" message, this one is flipping off the viewer with its middle finger.** The neon cactus glows in a light blue or teal color against a vibrant purple background, creating a bold contrast.

The cactus is sitting on a small purple platform, with a black cord extending from its base. The surrounding environment includes some purple, spiky objects that resemble rocks or perhaps more neon art. The overall aesthetic is very modern and edgy, combining the iconic cactus shape with a provocative gesture to create a memorable and eye-catching display.



Can you describe this image?

Molmo
The image depicts a woman with shoulder-length blonde hair sitting at a white kitchen table. She's wearing a red tank top and is in the process of cutting a slice of pizza on a red plate in front of her. There's a glass of water to her right and another plate with a pizza slice nearby.

The kitchen appears to be from the 1980s, with white cabinets and appliances. A large water cooler is visible on the left side of the image. The walls are white, and there's a clock on the wall above the cabinets.

The scene is well-lit, with a light fixture visible on the ceiling. The overall atmosphere is casual and domestic, capturing a moment of everyday life in a kitchen setting.

SCRAMBLE-Molmo
The image shows a woman sitting at a kitchen table. She has blonde hair and is wearing a red tank top. On the table in front of her is a plate with what appears to be pizza. The kitchen around her looks quite cluttered, with various items on the counters and shelves. There's a water cooler on the counter to her left, and the refrigerator is visible to her right. The overall scene gives the impression of a lived-in, busy kitchen.



Is the food in front of the woman or on her side?

Molmo
The food is in front of the woman. She is sitting at a table with a plate of food directly in front of her. The plate appears to contain a slice of pizza, and there's a glass of water to her right.

SCRAMBLE-Molmo
The food is in front of the woman. She has a plate of pizza in front of her on the table.



Is the food to the side of the woman?

Molmo
Yes, the food is to the side of the woman. The woman is sitting at a table with a plate of food in front of her, but there's another plate of food visible to her right. This suggests that the food she's eating is on her left side, while the additional plate is on her right.

SCRAMBLE-Molmo
Yes, the food is to the side of the woman. There's a plate of pizza to her left, and another plate of pizza to her right. The woman appears to be reaching for a slice of pizza on her left.



Figure 13. **Conversing with SCRAMBLE-Molmo (Winogrand example).** Both Molmo and SCRAMBLE-Molmo seem to read into the symbolism that the middle arm of the cactus looks like a middle finger and gives the image a rebellious tone.

Figure 14. **Conversing with SCRAMBLE-Molmo (EQBen example).** This is another examples from EQBen that SCRAMBLE-Molmo gets wrong but Molmo gets right. On asking to describe the image, both models do a reasonable job. On the 2nd question too both models respond that the food is in front of the woman. Somewhat contradictorily though, on the last question, both models answer yes.