Supplementary Materials for SCRAMBLe : Enhancing Multimodal LLM Compositionality with Synthetic Preference Data

Samarth Mishra Kate Saenko Venkatesh Saligrama
Boston University
{samarthm, saenko, srv}@bu.edu

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 [3] 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 [5] 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 [2] 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. 1. 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 [4].

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.

```
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 1. **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.

Results. The results of tuning a base LLaVA-1.5-13B

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
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	$\mathbf{38.6 \pm 0.1}$

Table 1. **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.

model on data generated with this approach are in Tab. 1 (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 examples 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. 2. 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 that SCRAMBLe's chain of thought approach.

B. Adversarial Refinement

In Sec. 3.2 of the main paper 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.

In Tab. 3 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.

Algorithm 1 Adversarial Refinement

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

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

C. Conversing with SCRAMBLe-Molmo : More Examples

More examples of conversive with Molmo and SCRAMBLe-Molmo are in Figs. 2 to 6. 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 (A100/A6000/A40/L40S/L40/RTX6000ada). We used the PyTorch framework [6] and our code for training MLLMs is based on Huggingface Transformers [9], TRL [8] and POVID [10]. We will upload our tuned models along with our synthetic data to Huggingface hub [1] 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 [5] and as the plausibility model we used Vera [3]. These models were used both for filtering

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.	A cart is being pushed by a white horse up a street.	A white horse pushing a cart down a street.
Close-up of bins of food that include broccoli and bread.	(Att) NA	Close-up of bins of food that exclude broccoli and include bread.	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 2. **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 3. **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.

as well as feedback generation.

D.2. Training

We used the direct preference optimization(DPO) [7] objective for preference tuning, as described in Sec. 3.3 of the main paper. 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 (this was selected from among candidate values $\{8, 16, 32, 64\}$ by validation performance over SugarCREPE-swap set). 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 of the main paper 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. 4 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-

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 4. 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.

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.
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.
```

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
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
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.
```

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"
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:
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'.
```

Figure 2. **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.

References

[1] 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. 2

[2] Matthew Honnibal and Ines Montani. spacy: Industrial-strength natural language processing in python, 2015. 1

[3] 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

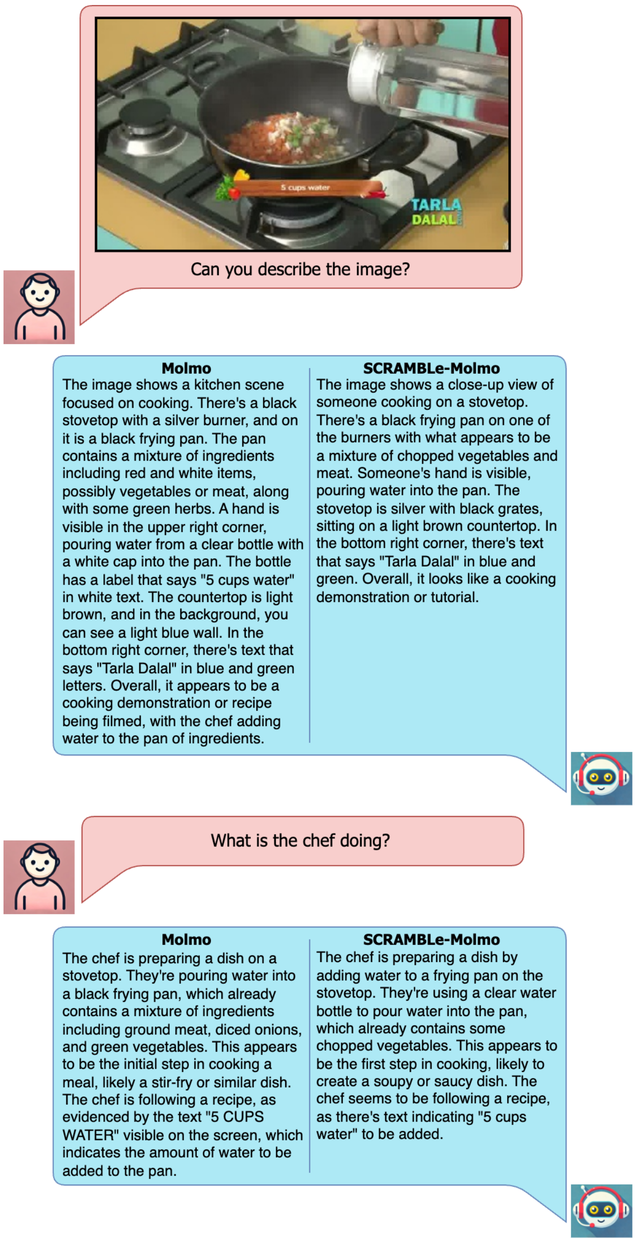

Figure 3. **Conversing with SCRAMBLe-Molmo (EQBen example).** The image is from an EQBen example that Molmo gets right but SCRAMBLe-Molmo does not. However, in conversation, SCRAMBLe-Molmo correctly identifies all elements of the scene and answers the question no worse than Molmo.

Linguistics. 1, 2

[4] 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. 1

[5] 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. 1, 2

[6] 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. 2

[7] 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

[8] 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. 2

[9] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, 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. 2

[10] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning, 2024. 2

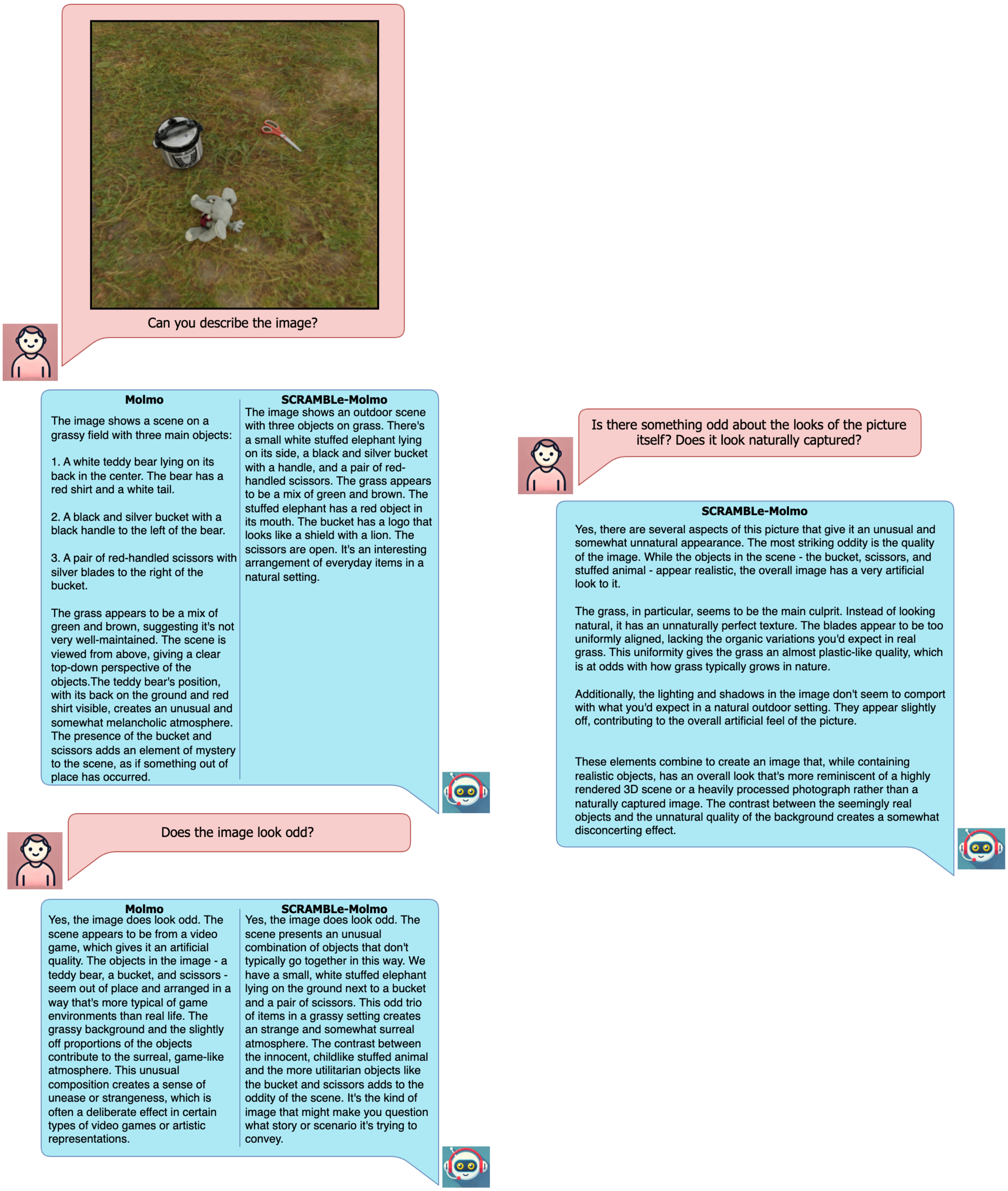

Figure 4. **Conversing with SCRAMBLe-Molmo (EQBen example).** This is another EQBen example that Molmo gets right but SCRAMBLe-Molmo does not. SCRAMBLe-Molmo does correctly identify the elements of the image. As a further question, when asked if there is something odd about the image, Molmo directly 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.

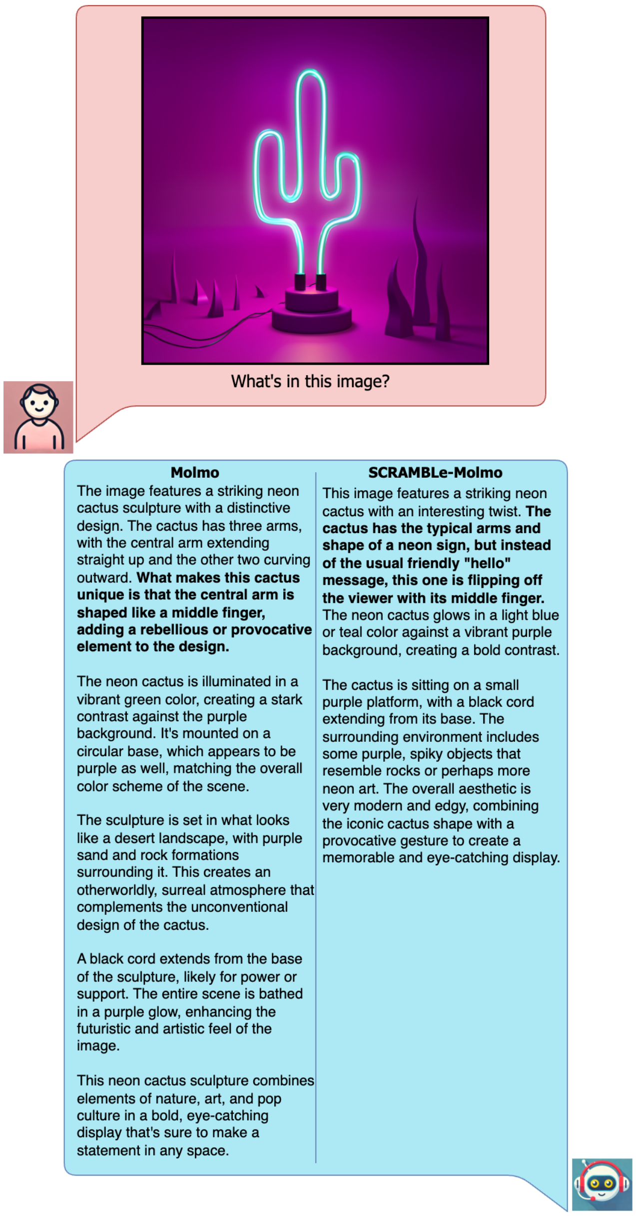

Figure 5. **Conversing with SCRAMBLe-Molmo (Winoground example).** Both Molmo and SCRAMBLe-Molmo seem to read into the symbolism that the middle arm of the cactus looks like a middle finger and gives the image a rebellious tone.

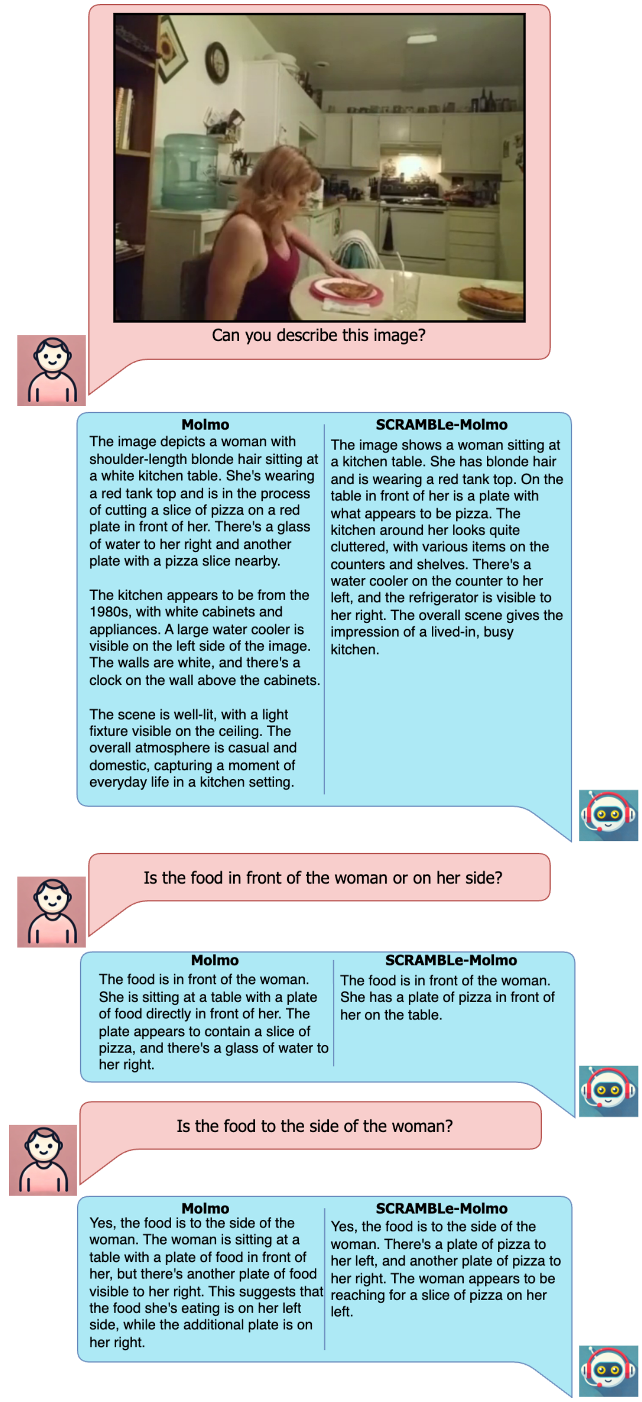

Figure 6. **Conversing with SCRAMBLe-Molmo (EQBen example).** This is another examples from EQBen that SCRAMBLe-Molmo gets wrong but Molmo gets right. On asking to describe the image, both models do a reasonable job. On the 2nd question too both models respond that the food is in front of the woman. Somewhat contradictorily though, on the last question, both models answer yes.