Supplementary Material for "BEAF: Observing BEfore-AFter Changes to Evaluate Hallucination in Vision-language Models"

Moon Ye-Bin¹*⁽⁰⁾ Nam Hyeon-Woo¹*⁽⁰⁾ Wonseok Choi²⁽⁰⁾ Tae-Hyun Oh^{1,2,3}⁽⁰⁾

¹Dept. of EE, POSTECH, Korea ²Grad. School of AI, POSTECH, Korea ³Institute for Convergence Research and Education in Advanced Technology, Yonsei University, Korea. {ybmoon, hyeonw.nam, wonseok.c, taehyun}@postech.ac.kr https://beafbench.github.io/

In this supplementary material, we include additional discussion about our metric, the filtering criteria used during stage 2 in the image manipulation pipeline, along with additional results, which are not included in the main paper.

A Comparison of TU and Accuracy

We propose four new change-aware metrics in Sec. 3.2 in the main paper. True Understanding (TU) measures whether the models correctly change their answers about manipulated objects. IGnorance (IG) measures the extent to which models lack knowledge about specific scene information. StuBbornness (SB) measures the extent to which models adhere to their initial answers. InDecision (ID) measures whether the models change their answers about unmanipulated objects. The evaluation results in Sec. 4 in the main paper demonstrate that by tracking the changes in answers along the vision axis (original and manipulated images), we can evaluate hallucination in VLMs more accurately.

More complex scenes make VLMs more confused like humans. In this section, we compare TU and accuracy (ACC) in terms of image complexity to demonstrate that our proposed TU is better calibrated than the traditional metric of accuracy. Motivation of this analysis comes from our postulation that the more complex the scene is, the easier it is for VLMs to be confused like humans; humans tend to be likely confused to miss some objects if there are many objects in the scene. In Fig. S1, we measure our TU and ACC along the image complexity. We represent image complexity in two ways: (1) image feature distance and (2) the number of objects in a scene. The image features are extracted from DINOv2 [5], and we measure the distances between global (class token) and the rest of the features. We measure the standard deviation of the distances and use it as the image complexity value. The high standard deviation means that many local features (local token) are far from the global feature (class token); thus the image contains diverse features, *i.e.*, a complex scene. To count the number of objects in an original image scene, we use Recognize Anything Model (RAM) [8].

^{*} Authors contributed equally to this work.





Fig. S1: Comparison of accuracy (ACC) and TU along the image complexity. We represent image complexity as a value using (1) image feature distance and (2) the number of objects in a scene. These image complexity values are measured based on the corresponding original image. TU and ACC are measured for Shikra-7B. The slope of the trend line for TU is steeper than for ACC. For the graph according to the number of object (left), the slopes of the trend lines are -0.209 and -0.428 (for simplicity, we consider each x-axis interval to be 1.), respectively, *i.e.*, decreasing twice faster. Assuming that our TU metric is a more calibrated metric than ACC, it clearly shows that the more complex a given image is, the more likely the model gets confused to produce hallucinatory answers.

In the first row of Fig. S1, the previous measure, accuracy (ACC), is slightly decreased as the image complexity increases. The results may imply that VLMs are susceptible to complex scenes, but it is hard to conclude because the significance of the trend may be low within the deviation ranges of data. Comparing ACC (first row) and TU (last row), the trend line slope of TU is clearly steeper than that of ACC. Our metric shows the relationship between image difficulties and model hallucinations more distinguishably than ACC, indicating that our TU is a more calibrated metric than ACC.

B Filtering Criteria

Our BEfore-AFter (BEAF) dataset comprises original and manipulated images. We manipulate each original image with the three-stage pipeline to obtain the manipulated images (as shown in Fig. 3 in the main paper). In stage 2, human annotators filter the automatically manipulated images based on the filtering criteria. These manipulated images are either directly used as data, passed to the next stage, or discarded. In Fig. S2, we visualize the filtered samples that are either passed to the next stage or discarded, along with the corresponding filtering criterion applied.

Mask error. In the first row on the left side, although the carrots are automatically removed during the first stage, some carrots remain. We categorize it as a mask error because the masks fail to cover all carrot regions. Consequently, this manipulated image is passed to the human-guided manipulation stage, where the carrots are ultimately removed. In the second row, a person is removed, but the borders of the arm remain. Such remaining parts may indicate the presence of a person in the image, leading us to categorize this example as a mask error and pass it to the next stage for further processing.

Shadow. In the third and last row on the left side, the target objects (a person and a motorcycle, respectively) are well removed. Still, their shadows remain in the manipulated image. Shadows can provide clues about the existence of the removed objects; therefore, we classify these cases as shadow errors and pass them to the next stage for shadow removal.

Object size. In the first row on the right side, the size of the sofa is too large to be completely removed. Even after erasing the sofa, traces remain visible, giving the impression that the dog is still on the sofa. This phenomenon occurs when large objects are erased, regardless of whether the manipulation is done by a human or automatically. Consequently, we decided not to use such images.

Lack of scene information. In the last two rows on the right side, there is a lack of surrounding scenes to fill the place left by the removed objects. This criterion is also closely related to the object size criterion, *i.e.*, there could be a lack of scene information if the object covers a significant portion of the image. Images falling under this category were also deemed unsuitable for use because the quality does not improve even with direct human manipulation. In addition to the aforementioned criteria, there are other factors, such as severe occlusion.

C Additional Results

In this section, we include additional samples and qualitative results, including correctness heatmap, CLIP Score, and open-generation. We also include classification results to check whether the removed object is not detected.

Image samples in our BEAF dataset. In Fig. S3, we present samples from our BEAF dataset, showcasing both original and manipulated images. The manipulated images demonstrate the effective removal of objects belonging to the target semantic class from the original images. Note that our BEAF dataset contains both original and manipulated images.

Additional visualization of image-wise object relationship results. In addition to Fig. 4 in the main paper, we visualize the object relationship for other images in Fig. S4. The x-axis is the text axis where each object is associated

4 Ye-Bin et al.



Fig. S2: Filtering criteria and corresponding samples. During stage 2 of the image manipulation pipeline, we filter the automatically manipulated images based on the predefined criterion. The manipulated images with mask error or shadow are passed to stage 3 (left); others are unused (right). We visualize the samples with the corresponding filtering criterion.

with the question, "Is there {object}?"; the y-axis is the vision axis where the object is removed from the original image. For example, in the first sample, the coordinate of (cat, clock) corresponds with the correctness of the answer to the question "Is there a cat?" when the image lacks the clock. The blue color indicates correctness, while the red indicates incorrectness.

Bottom 10 CLIP Score. While Fig. 6 in the main paper visualizes samples with CLIP Score and their differences between original and manipulated images with high values, Fig. S5 additionally visualize samples with the CLIP score, focusing on the lowest 10 differences. Notably, the semantic classes of the removed object from the samples in Fig. 6 are diverse, whereas the ones in Fig. S5 are mostly "person." However, the detector (Faster R-CNN [6]) still detects the person in the original images but fails to do so in the manipulated images, con-

sistent with the result of the samples with a high CLIP score difference. This suggests that CLIP often has hallucinations for the semantic class "person."

Classification. We adopt the pre-trained model of removal to manipulate the original scene. Although we filter out the low-quality images and observe whether the objects are well removed, there remains a question regarding the completeness of removal from the machine's perspective. To measure this, we define a concept of agreement wherein the

removed object should not be detected from the manipulated image. First, we remove the object from the original image. Next, we apply the detector to both the original and manipulated images. Then, we count if the object removed from the

Table	e S1: Cla	ssifi	cation	result	. We
check	whether	${\rm the}$	remove	d obje	ct is
not de	etected us	sing t	he obje	ct dete	ctor.

Model	Agreement	
Faster R-CNN	95.96	

original image is undetected in the manipulated images. The result, illustrated in S1, indicates that 95.96% of the removed objects are undetected from the manipulated images.

Additional open generation results. In addition to Fig. 7 in the main paper, we visualize the examples of the open-ended generation task in Fig. S6-S9. The results show that vision language models (VLMs), such as LLaVA-v1.5 [4], InstructBLIP [3], Shikra [2], mPLUG-Owl2 [7], GPT4 [1], tend to hallucinate in the open-ended generation answers. When we ask "Describe this image.", they often mention the removed object or other objects correlated but not grounded on the given image. It is worth noting that GPT4 is relatively robust to hallucination compared to the other VLMs.

References

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv (2023) 5, 10
- Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv (2023) 5
- Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
 5
- 4. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv (2023) 5
- Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023) 1
- Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NeurIPS) (2015) 4, 9

- 6 Ye-Bin et al.
- Ye, Q., Xu, H., Ye, J., Yan, M., Liu, H., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. arXiv (2023) 5
- Zhang, Y., Huang, X., Ma, J., Li, Z., Luo, Z., Xie, Y., Qin, Y., Luo, T., Li, Y., Liu, S., et al.: Recognize anything: A strong image tagging model. arXiv (2023) 1



Fig. S3: Samples of the original and manipulated images in the BEAF dataset. We visualize sample images in our BEAF dataset. The first column contains original images, and the rest of the columns contain manipulated images. The removed object is noted below each image.

8 Ye-Bin et al.



Fig. S4: Visualization of image-wise object relationship. We visualize the original images and the corresponding object relationship with the text (x-axis) and vision (y-axis) axes. We color the correct answer blue and the wrong one red. For example, a heatmap entirely in blue indicates that the answers to all questions are correct. On the text axis, the objects represent the questions posed, whereas, on the vision axis, they correspond to the objects manipulated in the image. The "none" on the vision axis means the original image, *i.e.*, none of the objects is removed. This approach allows us to analyze the influences between objects within a scene simultaneously.



Fig. S5: Bottom 10 samples according to the CLIP Score. The magnifying glass icon represents whether Faster R-CNN [6] detects the removed object. The CLIP Score of each sample is displayed below the image. We measure the CLIP score between an object and a prompt "a photo of {object}." Most of the bottom 10 samples have removed "person" from the image.

10 Ye-Bin et al.





The image features a yellow and green trolley car traveling down a street, passing by a building. The trolley car is positioned in the middle of the scene, with its front facing the viewer. There are multiple traffic lights visible in the image, with one on the left side, another on the right side, and a third one further back. In addition to the trolley car, there are two people in the scene, one standing closer to the left side and the other near the right side of the image. A fire hydrant can also be seen on the left side of the street, close to the traffic lights....

InstructBLIP

The image depicts a yellow and green trolley car traveling down a city street. The trolley is surrounded by various vehicles, including cars, trucks, and buses. There are several traffic lights visible in the scene, indicating that the trolley is navigating through a busy urban area. The trolley appears to be moving at a moderate pace, possibly on its way to pick up passengers or transport them around the city. In addition to the trolley, there are several pedestrians walking along the sidewalks, adding to the bustling atmosphere of the scene.



The image features a yellow and green train car traveling down a city street. It appears to be a cable car, and it is moving along the tracks beside a sidewalk. There are several traffic lights located at various points along the street, indicating that the train is passing through an intersection. In addition to the train and traffic lights, there are a few people visible along the sidewalk. Jikely pedestrians walking by or waiting at the sidewalk. A bench can also be seen in the scene, providing a place for people to sit while waiting for public transportation.



The image features a yellow and green trolley car traveling down a street, passing under a traffic light. The trolley is positioned in the middle of the scene, with the traffic light located above it. There are several traffic lights in the area, with one on the left side of the trolley, another on the right side, and a third one further back on the right side of the scene. In addition to the trolley, there are two people visible in the image. One person is standing near the left side of the trolley, while the other person is located closer to the right side of the scene.



This image features a vibrant street scene with a vintage yellow and green trolley car, marked "MATA," at the forefront. It appears to be an electric streetcar, connected to overhead wires, likely indicating a functioning public transport system. The streetcar is on a street with visible traffic lights, one of which is in the amber phase. There are also "No Left Turn" signs displayed on the traffic lights. In the background, there's a building with a sign that reads "The State Bank of [partially obscured]". The sky is clear, suggesting a pleasant day. The environment suggests this might be a historical or tourist area, possibly preserving the heritage of streetcar transportation. Manipulated



The image features a yellow and green trolley car traveling down a street, with a building visible in the background. The trolley car is positioned in the middle of the scene, and it appears to be moving along the tracks. There are several people in the image, with one person standing near the left side of the trolley car, another person closer to the right side, and a third person further back on the right side of the scene. Additionally, there is a traffic light visible in the background, indicating that the trolley car is operating in an urban environment.

The image depicts a yellow and green trolley car traveling down a city street. The trolley is passing by a building on the left side of the street, and there are several people visible in the scene. One person is standing near the front of the trolley, while others are scattered throughout the scene. There is also a stop sign located on the right side of the street, indicating that the trolley is coming to a halt at the next stop. The trolley appears to be part of a public transportation system, providing a convenient way for people to travel around the city.

The image features a yellow and green train moving down the street, possibly a trolley car or tran. It is traveling on a track that runs beneast the street level. The train appears to be passing through a town, with a building visible in the background. There are several people in the scene, some standing mear the train and others scattered along the street. One person is wearing a tie, adding a touch of formality to the casual setting. Additionally, a traffic light can be seen on the right side of the image, further indicating that the train is part of the city's transportation system.

The image features a yellow and green trolley car traveling down a street, passing by a building. The trolley is positioned in the middle of the scene, with a few other cars visible in the background. The street appears to be a busy area, with multiple traffic lights scattered throughout the scene. There are several people present in the image, some of them standing near the trolley and others walking along the sidewalk. A handbag can be seen on the ground near one of the pedestrians. The overall atmosphere of the scene suggests a bustling urban environment with public transportation and pedestrians going about their daily activities.

This image features a vintage green and yellow streetcar, marked with "MATA", traveling on a city street. Overhead, there are electric cables that appear to be powering the streetcar, with a couple of those cables connected to the streetcar's roof via a trolley pole. On the side of the road, there's a sign for "the State Bank of Memphis". The environment suggests a sunny day with a clear blue sky, and there are some green trees visible. The setting appears to be in an urban area with a mix of modern and possibly historical elements, hinting at a city that values both its past and present.

Fig. S6: Additional qualitative example (1) on open-ended generation task. We report the changes in answers for the open-ended generation task with the prompt "Describe this image." as the image undergoes manipulation. The red color high-lights the hallucination parts. We additionally report the GPT4 [1] results for reference. In this example, the traffic lights are removed.





6

The image features a yellow train parked at a train station, with a person standing on the platform next to it. The train occupies a significant portion of the scene, stretching from the left side to the right side of the image. There are several other people in the scene, some of them standing near the train and others further away. A handbag can be seen on the platform, likely belonging to one of the passengers. The train station appears to be a busy place, with people waiting for their train or just passing through.

The image depicts a busy train station with a yellow passenger train parked on the tracks. There are several people standing near the train, waiting to board or disembark. Some of them are positioned closer to the front of the train, while others are further back. In addition to the people, there are several handbags scattered throughout the scene, likely belonging to the passengers waiting for their transportation. The overall atmosphere is lively and busting, reflecting the activity at the train station.



The image features a yellow train traveling down the tracks, pulling into a station. The train appears to be a part of a commuter train service, and it is stopped at a platform. Several people are waiting on the platform, likely preparing to board the train or waiting for someone who will be getting off. There are various passengers visible on the platform, with some standing closer to each other and others spaced further apart. The passengers seem to be engaged in different activities or conversations as they wait for the train to come to a complete stop.

mPLUG-Owl2

The image features a yellow train stopped at a train station, with several people waiting on the platform. There are at least 11 people visible in the scene, some closer to the train and others further away. The train occupies a significant portion of the image, extending from the left side to the right side of the platform. In addition to the train and people, there are a few other objects in the scene. A handbag can be seen near the center of the platform, and a backpack is located closer to the right side. A bench is also present, situated near the right edge of the platform.



The image depicts a bright yellow train stopped at a railway station. It's a sunny day, and the train, with a classic design, has multiple cars. The station has an overhead cover, and there is a platform where passengers are waiting. Above the train, there is an electric line structure, suggesting that this is an electric train. Signs and structures around the platform have a design that indicates the photo might have been taken in Japan, as suggested by the script visible on the signage. There stands, while one person appears to be sitting on a bench. The environment suggests a suburban or possibly semi-urban train station.

Manipulated



The image features a train station with a train on the tracks, waiting for passengers to board. A woman is standing near the train, possibly waiting to board or just disembarking. There are several other people in the scene, some of them closer to the train and others further away. The train station has a platform with benches for passengers to sit and wait. There are two benches visible in the scene, one near the center and another towards the right side. The platform is also equipped with a clock, allowing passengers to keep track of time.

The image depicts a busy train station with several people waiting on the platform. There are at least 10 people standing on the platform, some closer to the tracks and others further away. A train can be seen approaching the station from the left side of the image, while another train is visible on the right side of the platform. In addition to the trains, there are two benches located on the platform, one near the middle of the scene and the other closer to the left side of the image. The benches provide seating options for passengers waiting for their next train. Overall, the scene captures the hustle and bustle of a busy train station during the day.

The image features a train station with a train pulled into it, stopped at a platform. The platform has a covering, and several people are waiting on it, possibly getting ready to board the train or just waiting for its arrival. In total, there are 13 people visible in the scene, standing in various positions near the train. Two backpacks can be seen among the passengers, one near the left side of the platform and the other closer to the center. A couple of handbags are also present, one near the left edge of the platform, and the other closer to the right side. The passengers appear to be engaged in their own activities or conversations while they wait for the train.

The image depicts a train station with a group of people standing on the platform, waiting for their train. There are at least ten people visible in the scene, with some standing closer to the train tracks and others further away. The train station features a train on the tracks, occupying a significant portion of the image. The platform is surrounded by a few benches, providing seating for the waiting passengers. A handbag can be seen placed on the ground near one of the people, likely belonging to one of the passengers.

This image shows a railway platform with people waiting. On the right, there are several passengers standing or sitting, some engaged with their phones. In the foreground, there's a train track leading away to the left of the frame, where the lines converge at a point due to perspective. Overhead, there are electrical wires, suggesting this is an electrified rail system. The platform is partially shaded by a roof structure, and there's a building with more stories in the background, indicating an urban environment. There's a digital signboard, likely indicating train schedules, hanging from the roof of the platform. The weather seems clear, as it is brightly lit with sunlight.

Fig. S7: Additional qualitative example (2) on open-ended generation task. In this example, the train is removed.

12Ye-Bin et al.





The image showcases a large, clean, and well-organized kitchen with wooden cabinets and a tile floor. The kitchen is equipped with a stainless steel refrigerator, a microwave, and an oven. The refrigerator is located on the left side of the kitchen, while the microwave is placed above the oven, which is situated on the right side. There are two sinks in the kitchen, one located near the center and the other towards the left side. A bottle can be seen on the countertop, and a bool is placed on the left side of the kitchen. The overall atmosphere of the kitchen is bright and inviting, making it an ideal space for cooking and entertaining.



The image features a newly remodeled kitchen with stainless steel appliances, including a dishwasher, oven, microwave, and refrigerator. The kitchen is equipped with wooden cobinets and countertops, giving it a warm and inviting atmosphere. There are several bottles scattered throughout the kitchen, likely used for cooking or food preparation. Additionally, there i a bodk placed on the countertop, possibly containing ingredients for a meal or snack. Overall, the kitchen appears to be well-maintained and ready for use. is



The image shows a spacious kitchen with a variety of appliances and furniture. The kitchen features a metallic stove top oven placed next to a microwave oven, both positioned above a sink. The contertops appear clean, and there are no people in the kitchen. Several kitchen items can be seen on the countertops, such as a boxl, a knife, and a spoon. There are also three bottles placed near each other on the countertop, possibly containing cooking ingredients or kitchen supplies. The kitchen is equipped with a refrigerator, which can be seen in the far right corner of the room. Overall, the kitchen appears clean and well-organized.

22 mPLUG-Owl2 The image features a large, empty kitchen with wooden cabinets and a stainless steel refrigerator. The kitchen is Clean and well-organized, with a sink positioned towards the left side of the room. A microwave is mounted above the sink, and a dining table is located in the middle of the room. There are several bottles placed around the kitchen, with some near the sink and others scattered around the countertops. A cup can also be seen on the countertop, close to the sink. The overall atmosphere of the kitchen is inviting and ready for someone to move in and start cooking.

This image features a modern kitchen with wooden cabinetry in a light brown finish. The appliances, including a refrigerator, dishwasher, and a gas range with an over-the-range microwave, are made of stainless steel. The countertops are black, which contrasts with the cabinetry and complements the appliances. There is a double-basin stainless steel sink with a window above it, providing natural light and a view outside. The floor is tiled in a light color that matches the overall neutral color scheme of the space. The walls are painted in a soft beige, and the room is well-lit, suggesting the photo was taken during the day.



The image showcases a large, clean, and well-organized kitchen with wooden cabinets and a black countertop. The kitchen features a sink, an oven, and a microwave, all placed in a functional layout. The oven is located on the right side of the kitchen, while the microwave is positioned above the countertop. There are several bottles placed on the countertop. There are several the sink and two others closer to the oven. A boal can also be seen on the countertop, adding to the kitchen's well-equiped appearance. The overal latmosphere of the kitchen is inviting and ready for use.

The image depicts a spacious and well-appointed kitchen in a newly constructed home. The kitchen features wooden cabinets, a black countertop, and a stainless steel slnk. There is a microwave oven mounted on the wall above the slnk. A window can be seen in the background, providing natural light to the room. In addition to the kitchen appliances, there are several bottles scattered on the countertop, while others are located on the floor or near the sink. Overall, the kitchen appears to be well-equipped and ready for cooking and preparing meals.

The image features a spacious kitchen with a cente The image features a spacious kitchen with a center island that has a sink on it. The kitchen is furnished with wooden cabinets and has a tiled floor. There is a refrigerator positioned on the right side of the room, and a microwave oven placed above the stove top oven located in the center of the kitchen. Additionally, there are several i tems on the counter, including two bottles near the left side of the island and a bowl further to the left. A book can also be seen on the counter, possibly a cookbook or reference material for cookine. cooking.

The image features a large, empty kitchen with wooden cabinets and a black countertop. The kitchen is clean and well-organized, with a microave placed above the countertop and a sink positioned nearby. The cabinets are arranged in a way that maximizes the available space, creating a functional and efficient layout. In addition to the main kitchen appliances, there are a few items scattered around the room. A cup can be seen on the countertop, and a bottle is placed near the sink. A **book is also visible on the countertop, possibly a recipe book or a cookbook.** The overall atmosphere of the kitchen is inviting and neady for someone to start kitchen is inviting and ready for someone to start cooking and enjoying their meals.

This image displays a neatly organized kitchen corner with natural wood cabinets and black countertops. It has a gas stove and an over-the-range microwave, both in stainless steel, matching the side-by-side refrigerator and the dishwasher tucked under the counter. A window sits above a stainless steel double sink, letting in natural light. The kitchen's floor is laid with large, square, beige tiles, and it transitions to a hardwood floor in the adjoining room. The wall paint appears to be a pale beige, creating a warm, welcoming atmosphere. There are a few items on the countertops, including a **yellow sponge** and a dish soap bottle by the sink, and what appears to be a canister of oats. The lighting suggests it is daytime, with sufficient natural light entering the room.

Fig. S8: Additional qualitative example (3) on open-ended generation task. In this example, the ovens are removed.



Fig. S9: Additional qualitative example (4) on open-ended generation task. In this example, the frisbee is removed.

Original







R

Manipulated