

Visual Fact Checker: Enabling High-Fidelity Detailed Caption Generation

Supplementary Material

A. More Details on Models and Prompts

The models used in VisualFactChecker and baselines are:

- **Caption Proposer:** *BLIP-2-OPT-2.7B*, *InstructBLIP-7B*, *LLaVA-1.5-13B*, *Kosmos-2*
- **LLMs:** *GPT-4-0613*, *Llama-2-70B-chat*
- **Detector:** *Grounding DINO*
- **VQA:** *LLaVA-1.5-13B*

Below are the prompts we used in VisualFactChecker for captioning 2d images and 3d objects.

VISUALFACTCHECKER PROMPTS (2D IMAGES)

Proposal (LLaVA-1.5 / Kosmos-2):
Describe this image in detail.

Verification step-1 (GPT-4 / Llama-2):

This is a hard problem. Carefully summarize in ONE detailed caption based on the following two captions by different (possibly incorrect) people describing the same scene. Be sure to describe everything, and avoid hallucination.

Verification step-2 (GPT-4 / Llama-2):

I want to use an object detector to check the correctness of an image caption obtained by an image caption model. Can you help to parse the caption below and list all objects that could be detected with an object detection model in the image? Please only list the object name and ignore the description. Please use singular for all listed objects.
Caption: {}.

Please concatenate them together with “. ” as separation.

Verification step-3 (Grounding DINO):

N/A (Grounding DINO examines candidate objects in the checklist above and determines their presence in the image.)

Captioning (GPT-4 / Llama-2):

Objective: parse and modify image captions using the results from an object detection model (may have hallucination).

I will put the detection results to you in the following format: [{"object": "detected object name", "number": "number of detected object (N)"}]. Please follow the following steps:

Instructions:

Parse the object in the caption. (Note: only parse and modify the object (not color, action, size, shape, or other descriptions))

1. If the object was detected by the detection model, keep everything including all descriptions. For instance, if the original caption is: “a black and white panda toy”, if the toy was detected, keep all content even though the “panda” and “black and white” are not detected. Keep all descriptions about color, shape, actions, etc.
2. If the subject object was not detected, remove only the object. Do NOT remove color, shape, action, text and other descriptions.
3. Only decrease the object number if the detected object number is smaller than the caption number.

This is a hard problem. Please minimize modifications of the caption, and list all changes made along with the reasoning.

—BEGIN Detection results: —

{}

—END Detection results—

—BEGIN Raw caption: —

{}

—END Raw caption—

Please give the output in the following format:

Modification:

Updated caption:

We find Llama-2 may encounter difficulties in the last 2D captioning step due to the complexity of the prompt. A workaround is to use a script to compare the detection results with the object list from step-2 and identify objects to be removed. Then, employ Llama-2 solely for removing these objects and summarizing the description.

VISUALFACTCHECKER PROMPTS (3D OBJECTS)

Proposal (LLaVA-1.5):

Please describe the details of the 3D object, the detailed description will be used for a text to 3d model to generate this 3D object. Please provide details of the shape, color of each part, avoid imagination and solve it step by step.

Proposal (InstructBLIP):

Describe the 3D object in detail, step by step.

Verification step-1 (GPT-4 / Llama-2):

This is a hard problem. Carefully summarize in ONE detailed caption based on the following two captions by different (possibly incorrect) people describing the same 3D object. The detailed caption will be used for a text to 3D model to generate this 3D object. Be sure to describe everything, and avoid hallucination.

Verification step-2 (GPT-4 / Llama-2):

I have a description of a 3D object, the detailed caption will be used for a text to 3d model to generate the same 3D object. Some part of the description may have some hallucination, so I want to use a VQA model to double check some key description, Please ask at most 5 most important and concrete questions that I need to double check to improve the fidelity of the description. Please focus on the factors that influence the final text to 3D model generation.

Raw Caption: {}

Please output the 5 questions in a python list.

Verification step-3 (LLaVA-1.5):

N/A (LLaVA-1.5 takes questions above and raw view image as input and give answers).

Single view Captioning (GPT-4 / Llama-2):

I have a description of a 3D object, the detailed caption will be used for a text to 3d model to generate the same 3D object. Some part of the description may have some hallucination, so I use a VQA model to double check some key description.

Here is the original description that may contain hallucination: {}

Here are the questions and answers from a VQA model: {}

Please correct the description based on the VQA. I want to use the description as a prompt for a text-to-3D generation model to generate the same 3D object.

Object Captioning (GPT-4 / Llama-2):

Given a set of descriptions about the same 3D object from different camera views, please distill these descriptions into one concise caption:

Camera View 1 description: {}

Camera View 2 description: {}

B. Details on Human and GPT-4V Evaluation

Fig. 8 shows the Amazon Mechanical Turk human evaluation web UI. For GPT-4V evaluation, inspired by DALLE-3 [3], we craft a single prompt for evaluating all captions for a given image using GPT-4V (*gpt-4-vision-preview*). The prompt is as follows.

GPT-4V EVALUATION PROMPT (2D IMAGES)

You are responsible for judging the quality of image captions generated by computer programs. You will be presented with an image and captions describing the image. One of the captions is a reference caption and you are responsible for comparing other captions against the reference caption, one by one. Please evaluate captions in the following two aspects:

1. Correctness. A good caption should correctly describe image content, including the major objects and the surrounding scene.
2. Detailedness. A good caption should contain rich details about the image content, for example, spatial location of objects, clothing and actions of people, color, counting of objects, texts in the image, etc.

Given the image and a reference caption, you need to make decisions as to whether or not the given captions are better or worse than the reference caption in describing the overall image.

This is a difficult task. You should first evaluate the reference caption for the given image. And then evaluate each caption in the same way and compare against the reference caption. You should think out loud about your eventual conclusion for each caption, enumerating reasons why each of the captions is better or worse compared with the reference caption. After thinking out loud, you should output either "Better" or "Worse" for each caption depending on whether you think the caption is a better description of the image than the reference caption.

Reference Caption: {}
Caption 1: {}
Caption 2: {}
Caption 3: {}
Caption 4: {}

You must keep your thinking out loud and respond in less than 500 words. Please summarize your judgment at the end using the following format:

Caption 1: Better or Worse
...
Caption 4: Better or Worse

We provide an example of GPT-4V's response, corresponding to Fig. 10 (d), where the reference caption is our proposed VisualFactChecker, and caption 1 to 4 are BLIP-2, InstructBLIP, Kosmos-2 and LLaVA-1.5, respectively.

GPT-4V EVALUATION RESPONSE (2D IMAGES)

To evaluate each caption, I will consider the aspects of correctness and detailedness against the reference caption.

Reference Caption: The reference caption accurately describes that a person is snowboarding and performing a jump while kite flying, mentioning the colors of the outfit and the snowboard. It also provides context for the backdrop and gives a sense of the snowboarder's abilities.

Caption 1: This caption introduces the term "kite surfing", which could be misleading because that term is typically associated with water rather than snow. It lacks the detail of the outfit and the specifics of the snowboarding action.

Caption 2: This caption provides a summarization close to the reference but omits the colors of the outfit and the snowboard. It does, however, accurately affirm the skill and control of the snowboarder and the backdrop of the scene.

Caption 3: This caption mentions objects that are not present in the image, such as other people in the scene. It retains the mention of the green outfit and emphasizes the height of the jump and the kite's position but suggests parasailing or kiteboarding, which might not be accurate terms for this activity.

Caption 4: The description of the snowboarder's action and outfit matches the image well, and it adds the color of the sky. It also introduces additional context by mentioning other spectators in the scene, but it does not mention the kite flying aspect, which is essential for the activity depicted.

Judgment:
Caption 1: Worse
Caption 2: Worse
Caption 3: Worse
Caption 4: Worse

Similarly, we design prompt for evaluating captions for 3D object, given two rendered views (front view and back view). The prompt is shown as follows.

GPT-4V EVALUATION PROMPT (3D OBJECTS)

You are responsible for judging the quality of captions generated by computer programs designed for describing 3d models. You will be presented with two rendered images of the same 3d model and two captions describing the 3d model. Please evaluate captions in the following aspects:

1. Correctness. A good caption should correctly describe the 3d object, including what the object is and the major properties of the object.
2. Detailedness. A good caption should contain rich details about the 3d object, for example, part, color, material, function, etc.

Given rendered images of the 3d model and two captions, you need to make decisions on which caption is better in describing the overall 3d object.

This is a difficult task. You should first evaluate two captions based on rendered views of the 3d object. You should think out loud about your eventual conclusion for each caption, enumerating reasons why each of the captions is better or worse compared with the other. After thinking out loud, you should output either "Caption 1" or "Caption 2" depending on which caption you think is a better description of the 3d model.

Caption 1: {}
Caption 2: {}

You must keep your thinking out loud and respond in less than 200 words. Please output your judgment at the end using the following format:

Better Caption: Caption 1 or Caption 2

We provide an example of GPT-4V's response for 3D captioning evaluation, corresponding to Fig. 12 (b), where caption 1 is Cap3D and caption 2 is our VisualFactChecker.

GPT-4V EVALUATION RESPONSE (3D OBJECTS)

To evaluate the captions, let's consider the criteria of correctness and detailedness.

Caption 1 is very basic, only stating that the object is a "3D model of a racing car". This is correct but lacks detail.

Caption 2, on the other hand, gives a much more comprehensive description. It addresses the color scheme ("white and red sports car, with a black and red stripe"), the decorative elements ("checkered pattern on the hood"), specific parts of the car ("spoiler on the back", "curved roof", "black and sporty tires"), and even the background which helps to highlight the object ("positioned on a neutral background"). This caption goes beyond simply identifying the object as a racing car; it describes its aesthetic and functional elements, which aligns with both correctness and detailedness.

While Caption 1 is correct, it is very brief and does not provide any detail. Caption 2 is also correct but provides a rich and detailed description that would give someone who cannot see the image a clear understanding of the 3D model's appearance and character.

Better Caption: Caption 2

C. More Qualitative Results

We show more results of image captioning methods and their DALLE-3 reconstructed images using different generated captions (COCO 2D images in Fig. 9, 10, 11; Objaverse 3D objects in Fig. 12). We show more comparison with GPT-4V captions using Llama-2 as the LLM in Fig. 13.

D. Following Complex Prompts

By leveraging the LLM, VisualFactChecker can follow complex prompts to write captions in different styles. Examples shown in Fig. 14.

Which image caption is better?

Choose the caption that better describes the image. A good caption should be **correct**, **detailed**, and **well-written**:

1. Correct: A good caption should correctly describe image content, including the major objects and the surrounding scene.
2. Detailed: A good caption contains rich details about the image content, for example, the spatial location of objects, clothing, actions of people, color, counting of objects, texts in the image, etc.
3. Well-written: A good caption should be clear, coherent, and grammatically correct.



CaptionA: In the image, there is a wedding cake sitting on a dining table. The cake is three-tiered and decorated with a plaid pattern. A man and a woman can be seen sitting at the table, possibly enjoying the wedding celebration. There are several wine glasses placed around the table, indicating that wine might be served during the event. Additionally, there are two bottles of wine visible in the scene, one on the left side of the table and the other on the right side.

CaptionB: The image prominently features a white wedding cake adorned with blue and green tartan ribbons, elegantly displayed on a dining table. The table is tastefully set with various items including wine glasses, a cup, and a knife, suggesting a wedding reception setting. In the background, a couple, presumably the bride and groom, are engaged in conversation. The bride is distinguished by a tiara, while the groom is dressed in a suit. Additional elements like a dining room window in the background contribute to the overall ambiance of the scene.

CaptionA

CaptionB

Figure 8. Amazon Mechanical Turk web user interface.

Input image	Captions	DALLE-3 text-to-image
<p>(a)</p> 	<p>BLIP-2: a street sign on a pole</p> <p>InstructBLIP: The image features a street pole with multiple street signs attached to it. One of the signs reads "Woodward Ave", while the other sign reads "Woodward St". The pole is situated in front of a tall building, which can be seen in the background. There are several trees scattered throughout the scene, including one on the left side of the image and another on the right side. In addition to the trees, there are several cars parked around the area, adding to the urban atmosphere.</p> <p>Kosmos-2: The image features a street sign on a pole, with a green street sign reading "Madsworth" and a black street sign post reading "Madsworth."</p> <p>LLaVA-1.5: The image features a street corner with a wooden pole holding up two street signs. The signs are green and white, indicating the names of the intersecting streets. The pole is located in front of a large building, which serves as a backdrop for the scene. There are several people in the area, with one person standing close to the pole and the street signs, and two others further away. Additionally, there is a car parked on the street, slightly to the right of the pole.</p> <p>VisualFactChecker (Ours): The image prominently features a street sign pole with street signs attached to it, including a green sign reading "Madsworth" and a black sign post also reading "Madsworth." The signs indicate the names of the intersecting streets. The pole is situated in front of a large building, which serves as the backdrop for the scene, enhancing the visibility of the street names and their locations.</p>	
<p>(b)</p> 	<p>BLIP-2: a banana tree with a large purple flower</p> <p>InstructBLIP: The image depicts a lush green banana tree with several bananas hanging from its branches. One of the bananas is located closer to the top of the tree, while others are scattered throughout the branches. There is also a purple flower growing on the tree, adding a touch of color to the scene. In the background, there is a power line visible, indicating that the banana tree is situated in an urban or suburban area. Overall, the image captures a vibrant and lively setting, showcasing the natural beauty of the banana tree and its surroundings.</p> <p>Kosmos-2: The image features a large banana tree with a large purple flower growing on it. The flower is situated at the top of the tree, and it is surrounded by green leaves. The tree is situated in a yard, and there are several other banana trees visible in the background. The yard is filled with various objects, including a car, a truck, and a bicycle. The car is located near the left side of the image, while the truck is located towards the right side. The bicycle is located further back in the scene, closer to the right edge of the image.</p> <p>LLaVA-1.5: The image features a large banana tree with a bunch of bananas hanging from its top. The bananas are still green, indicating that they are unripe. The tree is surrounded by a fence, providing a sense of enclosure to the scene. The bananas are spread out across the tree, with some hanging closer to the top and others further down. The tree appears to be healthy and thriving, providing a natural and lush backdrop to the image.</p> <p>VisualFactChecker (Ours): The image showcases a large banana tree with a bunch of unripe, green bananas hanging from its top and a large purple flower growing on it, both creating a visually appealing display. The tree is surrounded by green leaves and is enclosed by a fence, providing a sense of protection. The tree is situated in a yard filled with various objects. One other banana tree is visible in the background.</p>	
<p>(c)</p> 	<p>BLIP-2: a young boy standing in front of a bus</p> <p>InstructBLIP: The image features a young boy standing in front of a bus, wearing a green sweatshirt and a baseball cap. There are several other people in the scene, including a man standing to the left of the boy, another man standing to the right of the boy, and a third man standing further away from the boy. In addition to the people, there are two backpacks visible in the scene. One backpack is located closer to the boy, while the other is positioned further away from him.</p> <p>Kosmos-2: The image features a young boy wearing a hat and a green sweater standing in front of a bus. He appears to be looking at the camera, possibly smiling. In the background, there is another person standing, possibly a bus driver or a passenger. The bus is parked on the side of the road, and the boy is standing close to it.</p> <p>LLaVA-1.5: The image features a young boy wearing a hat and a green sweater, standing in front of a bus. He appears to be looking at the camera, possibly posing for a picture. Another person, possibly a man, is standing behind the boy, also looking at the camera. They seem to be the main subjects of the photo. In the background, there is a bus visible, which could be the mode of transportation for the people in the scene. The boy and the man are the only two individuals in the image, with the boy being the main focus.</p> <p>VisualFactChecker (Ours): The image depicts a young boy, donned in a hat and a green sweater, standing in front of a bus parked on the side of a road in a desert-like environment. The boy appears to be looking at the camera with an ambiguous expression that could be interpreted as either a smile or a somewhat sad look. Behind the boy, another person, possibly a bus driver or a passenger, is standing, although their features are not clearly discernible. The bus seems to serve as a mode of transportation for the people in the area.</p>	

Figure 9. Comparison of different 2D image captioning methods (1/3). The left column is the input image from COCO, the middle column shows the caption generated by different methods (BLIP-2, InstructBLIP, Kosmos-2, LLaVA-1.5, and our proposed VisualFactChecker). In the right column, we use DALLE-3 as a text-to-image model to reconstruct 2D images using different generated captions.



Figure 10. Comparison of different 2D image captioning methods – more examples (2/3).

Input image	Captions	DALLE-3 text-to-image
<p>(g)</p> 	<p>BLIP-2: a boy sitting on a bench</p> <p>InstructBLIP: The image depicts a young boy sitting on a bench in a park. He is engrossed in reading a book, likely enjoying his leisure time in the outdoors. In the background, there is a playground with a slide and other play equipment visible. The park appears to be well-maintained and inviting for children to play and have fun.</p> <p>Kosmos-2: The image features a young boy sitting on a bench in a park, reading a book. He is surrounded by various playground equipment, including a slide and a swing. The park is located in a residential area, and there are several cars parked nearby. The boy appears to be enjoying his time in the park, possibly taking a break from playing or reading.</p> <p>LLaVA-1.5: The image features a young boy sitting on a park bench, reading a book. He is positioned in the middle of the scene, with the bench extending across the lower part of the image. The boy appears to be enjoying his time outdoors, engrossed in his reading. The park setting is evident by the presence of a swing set in the background, along with a few cars parked nearby. There are also a couple of other people in the scene, one standing close to the boy and another further away. Additionally, there is a backpack placed on the ground near the bench, possibly belonging to the boy or another park visitor.</p> <p>VisualFactChecker (Ours): The image depicts a young boy, donned in a yellow jacket, engrossed in reading a book while sitting on a bench in a park. The park, situated in a residential area, features a playground with various equipment such as a swing set and a slide in the background, suggesting a lively and engaging environment for children. Several cars parked nearby indicate the proximity of the park to the residential area. The boy seems to be enjoying his time, possibly taking a break from playing.</p>	
<p>(h)</p> 	<p>BLIP-2: a dog and a horse are standing next to each other</p> <p>InstructBLIP: The image depicts a donkey, a dog, and a horse standing on a grassy field next to a body of water. The donkey is the largest animal in the scene, while the dog and horse are smaller in size. The three animals seem to be interacting with each other, possibly playing or exploring their surroundings. The dog is positioned towards the left side of the image, while the horse is closer to the right side. The donkey is situated in the middle of the grassy field, surrounded by the other two animals.</p> <p>Kosmos-2: The image features a white donkey and a small brown and white dog standing next to each other on a grassy field. The donkey is standing on its hind legs, looking at the dog, while the dog is standing next and looking at the donkey. The field is located near a body of water, adding to the serene atmosphere of the scene.</p> <p>LLaVA-1.5: The image features a dog and a horse standing next to each other in a grassy field. The dog is positioned to the right of the horse, and both animals appear to be looking at something in the distance. The dog is smaller in size compared to the horse, and they seem to be enjoying their time together in the field.</p> <p>VisualFactChecker (Ours): The image depicts a white donkey, standing on its three legs in a grassy field near a body of water, creating a serene atmosphere. A small brown and white dog is positioned to the right of the donkey, both animals seemingly engaged in looking at each other or something in the distance.</p>	
<p>(i)</p> 	<p>BLIP-2: a man holding a tennis racket and a tennis ball</p> <p>InstructBLIP: In the image, a man is standing on a tennis court, holding a tennis racket and preparing to serve the ball. He is wearing a white shirt and shorts, and appears to be focused on the game. There are several tennis balls scattered around the court, some closer to the player and others further away. The background features a fence and trees, adding to the outdoor atmosphere of the scene.</p> <p>Kosmos-2: The image features a man in a white shirt and hat, wearing sunglasses, swinging a tennis racket and preparing to hit a ball. He is standing on a tennis court, and the green fence is visible behind him. The man appears to be focused and ready to play a game of tennis.</p> <p>LLaVA-1.5: The image features a man standing on a tennis court, holding a tennis racket in his hand. He appears to be in the middle of a game, possibly preparing to serve the ball. The man is wearing a white shirt and white shorts, which are typical attire for playing tennis. The tennis court is surrounded by a fence, and there is a green tarp visible in the background. The man's tennis racket is positioned towards the right side of the image, and he seems focused and ready for the next play.</p> <p>VisualFactChecker (Ours): The image features a man, dressed in a white shirt and shorts, wearing a hat and sunglasses, standing on a tennis court with a green fence visible behind him. He is holding a tennis racket in his right hand, swinging it and preparing to hit a ball, possibly tracking its movement. The man appears focused and ready, suggesting he is in the middle of a tennis match.</p>	

Figure 11. Comparison of different 2D image captioning methods – more examples (3/3).



















Input 3D object	Captions	MVDream text-to-3D	GPT-4V analysis
(a) 	Cap3D: A 3D model of a futuristic motorcycle with orange wheels.		"... straightforward and accurate ... However, it lacks detail about other elements of the motorcycle ..."
	VisualFactChecker (Ours): A 3D model of a sleek, futuristic motorcycle with a streamlined and aerodynamic design, featuring a predominantly black and red color scheme, positioned on a white background for a clean, minimalist appearance. The motorcycle is equipped with two large, red-accented wheels and a black seat, adding to its modern and stylish appearance.		"... provides a comprehensive description that encapsulates more of what is visible in the image ... not only correctly identifies the object as a futuristic motorcycle but also includes details about its design, color, and parts."
(b) 	Cap3D: 3D model of a racing car.		"... very basic ... correct but lacks detail"
	VisualFactChecker (Ours): The 3D object is a sleek and aerodynamic white and red sports car, with a black and red stripe on the side and a number "4" prominently displayed. The car has a checkered pattern on the hood, a spoiler on the back and a curved roof. The tires are black and sporty, complementing the car's overall design. The car is positioned on a neutral background, which accentuates its predominantly white color scheme with red accents and black stripes.		"... a much more comprehensive description ... correct but provides a rich and detailed description that would give someone who cannot see the image a clear understanding of the 3D model ..."
(c) 	Cap3D: Wooden tripod floor lamp with a 3D model.		"... partially correct ... a vague description. It does not provide any details ..."
	VisualFactChecker (Ours): The 3D object is a vintage-style, small wooden tripod with three legs providing stability. Attached to the top of the tripod is a round, silver, metal lamp with a reflective surface, giving it the appearance of a mirror. The wooden legs of the tripod are brown, while the lamp/mirror has a silver appearance. This functional and decorative piece of furniture or decoration, with its simple yet effective design, is set against a clean, white background.		"... offers a more precise description ... not only correctly identifies the object as a tripod with a lamp but also includes details about the materials, colors, and function ..."
(d) 	Cap3D: A 3D-rendered small cake with a pumpkin on top, on a blue plate and napkin.		"... correctly identifies the object ... no mention ... lacks detail ..."
	VisualFactChecker (Ours): The 3D object is a visually appealing dessert featuring a cake with a pumpkin on top, and a bird perched next to the pumpkin. The cake, which is brown and white, is placed on a blue plate. The pumpkin, made of chocolate, is orange and protrudes from the top of the cake. The bird is brown in color. The bird, pumpkin, and cake are all part of the same 3D object, creating a unique and delicious-looking dessert.		"... correctly describes ... rich in detail ... more vivid description ... acknowledges all visible parts of the model and their characteristics ..."
(e) 	Cap3D: 3D model of a blue armored robot with horns.		"... succinct, but it lacks the descriptive richness ..."
	VisualFactChecker (Ours): A 3D model of a rust-covered, blue robot with a segmented body, appears to be part of a futuristic or science-fiction setting. The robot is standing upright with arms positioned to the side. The robot has a helmet on its head, which is also blue. It has a pair of intimidating horns on its body, resembling armor or protective gear. The overall appearance of the robot is old and worn, yet unique.		"... goes much further in detail ... also specifies ... which isn't contradicted by the image ... provides context and a more vivid description"
(f) 	Cap3D: 3D model of an antique metal sewing machine.		"... correct. However, it lacks detailed information about the sewing machine's design, the desk it is placed on, and the overall setup."
	VisualFactChecker (Ours): The 3D object is an antique, metal sewing machine with a unique design, placed on a small, old-fashioned desk with a black top and a metal frame, supported by four legs. The desk, which also serves as the table for the sewing machine, has a metal frame and a black metal top. The sewing machine, with its black metal body, is centrally positioned on the desk, serving as the focal point. The sewing machine has a pedal on the side for operation. The entire setup is placed against a white background, highlighting the intricate details of the desk and sewing machine.		"... a far more detailed description ... correctly identifies the object and provides rich details about its parts, color, material, function, and the context ..."

Figure 12. Comparison of different 3D object captioning methods. The left column shows the input 3D objects, the middle column shows the captions generated from Cap3D and our VisualFactChecker. In the right column, we use MVDream [29] as a text-to-3D model to reconstruct 3D objects using different 3D captions (generated by Cap3D [23] and ours). From the results, we can see that the reconstructed 3D objects using Cap3D captions are less similar to the input ones compared with ours, suggesting their captions may not contain sufficient information to reconstruct the 3D object or incorrectly describe the visual contents. 3D objects reconstructed using our captions are more similar to the inputs.

Input image

Captions

(a)



GPT-4V: The image shows a family moment on the beach, with warm sunlight illuminating the scene. A man is kneeling on the sand, engaging with two young children, a boy and a girl. The man and the children are focused on playing with sand toys, including buckets and shovels. They appear to be in the process of building a sandcastle. The girl, wearing a polka-dot coat, is to the man's left, while the boy, wearing a patterned sweater and denim pants, is on his right. It's a candid, joyful moment, capturing the essence of family leisure time at the beach. The background features the soft blur of a hill, emphasizing the focus on the family.

VisualFactChecker (Ours): A father and his two children are spending quality time together on a sandy beach, engaging in fun activities like building sandcastles and playing with toys. The father is kneeling down on the sand, surrounded by his children, who seem to be enjoying their time together. There are several sandcastles and toys scattered around the scene, indicating that the family has been having a great time playing and building together. The atmosphere is filled with joy and togetherness, as the family members enjoy each other's company on the beach.

(b)



GPT-4V: The image features a young girl with blonde hair styled in braids, wearing glasses, and holding a chalkboard sign that reads "BACK TO SCHOOL" with a smiley face drawn below the text. She's smiling, looking towards the camera, and appears to be happy about the return to school. The girl is wearing a striped sleeveless top and is positioned against a vibrant yellow background, which creates a cheerful and energetic atmosphere. The chalkboard and the girl's pose suggest a promotional or celebratory picture for the beginning of a new school year.

VisualFactChecker (Ours): A cheerful young girl with glasses is excitedly holding a chalkboard sign that reads "Back to School" in front of a bright yellow wall. She is smiling and posing for the camera, showcasing her enthusiasm for the upcoming school year. The girl is wearing a striped shirt and a pair of blue jeans, and her glasses have a subtle glint from the light. The chalkboard sign is neatly written in white chalk, with the words "Back to School" and a smiling face standing out clearly against the dark background. The overall atmosphere of the scene is one of excitement and anticipation for the new school year.

(c)



GPT-4V: The image features a miniature red pickup truck toy positioned in the foreground, with a focus on its details, highlighting the glossy finish and classic design. In the background, slightly out of focus, is a stack of books and a small blue house model, both creating a sense of depth and context for the scene. The background appears to be outdoors, as indicated by the green hues that suggest vegetation. The overall setting seems to be a staged arrangement, perhaps used for educational purposes or as a decorative display. The shallow depth of field used in this photograph emphasizes the toy truck while softly blurring the background elements.

VisualFactChecker (Ours): A red toy truck is sitting on a table in front of a blue toy house, giving the impression of a playful and creative setup. The truck appears to be positioned as if it is driving past the house, with its front bumper near the house's front door. To the left of the house, there is a stack of books, possibly suggesting a cozy reading nook or a place for a child to play and learn. Overall, the scene seems to be designed to evoke a sense of whimsy and imagination, perhaps for a child's play area or a collector's display.

Figure 13. Comparison of VisualFactChecker with GPT-4V. Our method can generate high-fidelity detailed captions that closely match GPT-4V's quality. To compare with GPT-4V, we use Llama-2 as our LLM instead of GPT-4 when generating captions for the above images.

Input image



Additional Prompts

—

...Describe the image in **less than 20 words**

...Describe the image in **less than 50 words**

...Describe the image and
do not mention the background

...Describe the image, **only mention the background and the overall feeling.**
Do not mention the foreground objects

...Describe the image, **starting with the background and then foreground objects,**
ending with an overall feeling

Captions

A happy little girl is standing in a green field, wearing a plaid shirt and holding onto a string of pink balloons. The balloons are floating in the air, creating a playful and joyful atmosphere. The girl appears to be enjoying the moment, with a smile on her face. In the background, there is a vast expanse of green grass, stretching out as far as the eye can see. The sky above is a bright blue, with a few white clouds scattered about. The overall mood of the scene is one of carefree happiness and joy, with the pink balloons adding a touch of whimsy and playfulness to the atmosphere.

Joyful girl with pink balloons smiling in a vast green field under a clear blue.

A joyful girl in a plaid shirt smiles as she holds pink balloons in a vast, green field under a blue sky dotted with clouds, embodying whimsical happiness.

A joyful young girl, clad in a plaid shirt, delights in the moment as she clutches a string of pink balloons. The balloons, buoyant and playful, dance in the air above her, enhancing her smile with a whimsical charm.

A vast, green meadow stretches endlessly, under a brilliant blue sky dotted with a handful of fluffy white clouds. **This idyllic scene exudes a sense of whimsical joy and carefree happiness, inviting a playful and cheerful mood.**

In the backdrop of this serene scene lies a boundless expanse of lush green grass under a vivid blue sky dotted with a smattering of white clouds. **At the heart of this idyllic setting stands a joyous little girl, clad in a plaid shirt, her hands clutching a string of buoyant pink balloons that dance in the air. This moment captures a snapshot of blissful joy and playful spirit, framed by nature's sprawling beauty.**

Figure 14. Following complex prompts. By leveraging an LLM to write the final caption, VisualFactChecker can follow complex instructions to write captions in various styles.