Spatial457: A Diagnostic Benchmark for 6D Spatial Reasoning of Large Multimodal Models

Supplementary Material

A. New programs for 6D spatial reasoning

As introduced in Sec. 3.3, we create two types of new questions: **6D spatial reasoning** and **collision prediction**.

The 6D spatial reasoning questions introduce the new program 3D Relate, an extension of Relate from Johnson et al. [19]. It takes two arguments: a direction (left, right, front, or behind) and an object ID specifying the target. Unlike Relate, which operates in a 2D plane from the camera's view, 3D Relate works in 3D space from the object's perspective, as described in Sec. 3.3.

For collision prediction, we introduce the new program Relate Collision. This operation expands on the spatial reasoning of 3D Relate by incorporating motion-based inference. It assesses whether the movement of a target object, specified with directions like forward or backward, will result in a collision with other objects in the scene. This program enables models to predict interactions and anticipate future states based on both 3D location and orientation, pushing the boundaries of spatial reasoning to include dynamic scenarios.

B. Examples of Benchmarks

In this section, we present additional examples from the Spatial 457 benchmark. Each example consists of an image accompanied by seven questions spanning five levels of difficulty, as shown in Fig. 8.

C. Prompts for LLM

In this section, we present the detailed prompts designed for evaluating the Large Language Models (LLMs) on the Spatial457 benchmark. Each level targets a specific skill set, ranging from identifying single objects to reasoning about spatial relationships and occlusions in 3D space.

The prompts are structured with clear instructions, encouraging the model to first describe the scene comprehensively before formulating the answer. This approach ensures that the reasoning process is explicit and easy for interpretable. The output format is standardized as a JSON object, enabling straightforward evaluation and comparison against expected results. Figures 9–15 provide detailed examples of the prompts used at each level:



L4-3D Pose

Q: What is the color of the car which faces to the right? A: Purple

L5-6D spatial Q: Is the number of small purple aeroplanes in front of the tiny aeroplane greater than the number of tiny cars behind the road bike? A: No



L4-3D Pose Q: Which direction the SUV is facing? A: Left L5-6D spatial

Q: How many things are either large red buses on the right side of the cruiser or aeroplanes on the right side of the large double bus? A: 2



L4-3D Pose Q: What is the size of the gray object that faces the vertical direction as the gray motorbike? A: Large L5-6D spatial

Q: Does the large aeroplane to the right side of the blue motorbike have the same color as the big aeroplane that is on the left side of the big gray articulated bus? A: No L1-Single object: Q: What is the shape of the small brown thing? A: Sedan

L2-Multiple object

Q: How many things are either large blue double buses or tiny road bikes? A: 1

L3-2D spatial: Q: Are there fewer bicycles right

of the purple jet than tiny gray utility bikes? A: No

L4-Occlusion

Q: What is the color of the object that is occluded by the jet? A: Brown

L5-Collision:

Q: There is a bus will be collided by the small wagon if it moving forward; what size is it? A: Small

L1-Single object:

Q: What is the color of the articulated bus ? A: Red

L2-Multiple object

Q: The other big object that is the same color as the articulated bus is what shape? A: SUV

L3-2D spatial: Q: How many things are either objects that are behind the cruiser or big cars that are to the left of the green object? A: 6

L4-Occlusion Q: What is the size of the red object that is occluded? A: Large

L5-Collision:

Q: What is the shape of the small thing that the red bus will collide with if it moves backward? A: Fighter

L1-Single object:

Q: What is the color of the jet? A: Purple

L2-Multiple object Q: What number of purple objects

are big jets or small trucks? A: 1 L3-2D spatial:

Q: Does the big blue thing that is

to the left of the purple aeroplane have the same shape as the tiny thing? A: No L4-Occlusion

Q: What is the shape of the thing that is occluded by the scooter? A: Jet

L5-Collision:

Q: What is the color of the object the scooter will collide with if it moves backward? A: Blue

Figure 8. Examples from the Spatial457 Benchmark. Each example includes an image and seven questions that test the model across five difficulty levels.

Prompts for Level 1 - Single Object questions

System:

You are an intelligent chatbot designed to answer questions based on an image. Your task is to analyze the images, identify attributes of the objects and then determine the answer to the question.

INSTRUCTIONS:

Step 1: Identify the objects mentioned in the question, noting each object's shape, color, size and their locations. Objects can have a shape (e.g., 'airliner'), size ('small' or 'large'), color (e.g., 'blue'). The possible colors are: gray, blue, purple, brown, green, cyan, red, and yellow.

Step 2: Use the identified details to form a brief reasoning process for answering the question. Based on this reasoning, provide your answer in one of the following formats: (1) a phrase chosen from the available options: Small, Large, Airliner, Dirtbike, Road bike, Tandem bike, Suv, Wagon, Scooter, Mountain bike, Minivan, Sedan, School bus, Fighter, Chopper, Double bus, Truck, Articulated bus, Cruiser, Jet, Utility bike, Regular bus, Biplane, Gray, Blue, Purple, Brown, Green, Cyan, Red, Yellow, Left, Right, Front, Back, (2) an integer between 0-10 for questions like 'How many' or 'What is the number of,' or (3) 'Yes' or 'No' for questions asking 'Is there.'

Output your response using this JSON format: { 'Reasoning': '<reasoning>', 'Answer': '<answer>'}

User:

What is the shape of the big yellow thing?

GPT-40:

{'Reasoning': 'The image contains a green utility bike with a red seat, a small white biplane, a large yellow sedan, and two purple tandem bikes. The large yellow object is a sedan.', 'Answer': 'Sedan'}

Figure 9. The prompts used for L1- Single object questions. In the instruction, we encourage the model to first describe all the objects in the scenes and the reasoning process of answer the question. We format the reply into a dictionary format, where the answer can be decoded and compared with the ground truth answer.

D. Extension to Real Image Setting

As described in Sec. 5, we extend the *L4-Pose* questions to the real-world image dataset SUN-RGBD [36]. The scene structure from SUN-RGBD [36] provides object categories and 3D bounding boxes. We calculate the rotation of bounding boxes along the z-axis as ground truth orientations and construct questions using the object category and orientation, naming the set *L4-Pose-Real*. The question generation template remains the same as that of the original *L4-Pose* questions.

To reduce ambiguity in object localization, we add a red bounding box around target objects in the images and reference them in the questions to ensure clarity in object identification within the scene. For example: "Which direction are the chairs in the red bounding box facing?"

Prompts for Level 2 - Multiple Objects questions

System:

You are an intelligent chatbot designed to answer questions based on an image. Your task is to analyze the images, **identify attributes of multiple objects** and then determine the answer to the question.

INSTRUCTIONS:

Step 1: Identify the objects mentioned in the question, noting each object's shape, color, size and their locations. Objects can have a shape (e.g., 'airliner'), size ('small' or 'large'), color (e.g., 'blue'). The possible colors are: gray, blue, purple, brown, green, cyan, red, and yellow.

Step 2: Use the identified details to form a brief reasoning process for answering the question. Based on this reasoning, provide your answer in one of the following formats: (1) a phrase chosen from the available options: Small, Large, Airliner, Dirtbike, Road bike, Tandem bike, Suv, Wagon, Scooter, Mountain bike, Minivan, Sedan, School bus, Fighter, Chopper, Double bus, Truck, Articulated bus, Cruiser, Jet, Utility bike, Regular bus, Biplane, Gray, Blue, Purple, Brown, Green, Cyan, Red, Yellow, Left, Right, Front, Back, (2) an integer between 0-10 for questions like 'How many' or 'What is the number of,' or (3) 'Yes' or 'No' for questions asking 'Is there.'

Output your response using this JSON format: 'Reasoning': '<reasoning>', 'Answer': '<answer>', 'Related Objects': ['shape': '<shape>', 'size': '<size>', 'color': '<color>', ...]

User:

Is there any other thing that is the same color as the small utility bike?

GPT-40:

{'Reasoning': 'The objects in the image are: a small gray SUV, a small gray sedan, a small gray school bus, a small red utility bike, a small red tandem bike, a small red road bike, a small red mountain bike, a small red fighter jet, and a small purple biplane. The small utility bike is red. There are other objects that are also red: the tandem bike, the road bike, the mountain bike, and the fighter jet.', 'Answer': 'Yes'}

Figure 10. The prompts used for L2- Multiple object questions. In the instruction, we encourage the model to first describe all the objects in the scenes and the reasoning process of answer the question. We format the reply into a dictionary format, where the answer can be decoded and compared with the ground truth answer.

Prompts for Level 3 - 2D Spatial questions

System:

You are an intelligent chatbot designed to answer questions based on an image. Your task is to analyze the images, **identify attributes of multiple objects and their spatial relationship from 2D projected camera view**, and then determine the answer to the question.

INSTRUCTIONS:

Step 1: Identify the objects mentioned in the question, noting each object's shape, color, size and their locations. Objects can have a shape (e.g., 'airliner'), size ('small' or 'large'), color (e.g., 'blue'). The possible colors are: gray, blue, purple, brown, green, cyan, red, and yellow.

Step 2: Use the identified details to form a brief reasoning process for answering the question. Based on this reasoning, provide your answer in one of the following formats: (1) a phrase chosen from the available options: Small, Large, Airliner, Dirtbike, Road bike, Tandem bike, Suv, Wagon, Scooter, Mountain bike, Minivan, Sedan, School bus, Fighter, Chopper, Double bus, Truck, Articulated bus, Cruiser, Jet, Utility bike, Regular bus, Biplane, Gray, Blue, Purple, Brown, Green, Cyan, Red, Yellow, Left, Right, Front, Back, (2) an integer between 0-10 for questions like 'How many' or 'What is the number of,' or (3) 'Yes' or 'No' for questions asking 'Is there.'

Output your response using this JSON format: 'Reasoning': '<reasoning>', 'Answer': '<answer>', 'Related Objects': ['shape': '<shape>', 'size': '<size>', 'color': '<color>', ...]

User:

Is there a gray mountain bike that is to the left of the gray thing that is behind the tiny chopper?

GPT-40:

{'Reasoning': 'In the image, there are several objects including a gray car, a red car, a white bus, a purple airplane, and several bicycles. The gray car is located towards the left side of the image. To the right of the gray car, there is a red car. There are no brown bicycles to the right of the red car.', 'Answer': 'No'}

Figure 11. The prompts used for L3- 2D Spatial questions. In the instruction, we encourage the model to first describe all the objects in the scenes and the reasoning process of answer the question. We format the reply into a dictionary format, where the answer can be decoded and compared with the ground truth answer.

Prompts for Level 4 - Occlusion questions

System:

You are an intelligent chatbot designed to answer questions based on an image. Your task is to analyze the images, **identify attributes of multiple objects and their occlusions relationships**, and then determine the answer to the question.

INSTRUCTIONS:

Step 1: Identify the objects mentioned in the question, noting each object's shape, color, size and their locations. Objects can have a shape (e.g., 'airliner'), size ('small' or 'large'), color (e.g., 'blue'). The possible colors are: gray, blue, purple, brown, green, cyan, red, and yellow.

Step 2: Use the identified details to form a brief reasoning process for answering the question. Based on this reasoning, provide your answer in one of the following formats: (1) a phrase chosen from the available options: Small, Large, Airliner, Dirtbike, Road bike, Tandem bike, Suv, Wagon, Scooter, Mountain bike, Minivan, Sedan, School bus, Fighter, Chopper, Double bus, Truck, Articulated bus, Cruiser, Jet, Utility bike, Regular bus, Biplane, Gray, Blue, Purple, Brown, Green, Cyan, Red, Yellow, Left, Right, Front, Back, (2) an integer between 0-10 for questions like 'How many' or 'What is the number of,' or (3) 'Yes' or 'No' for questions asking 'Is there.'

Output your response using this JSON format: 'Reasoning': '<reasoning>', 'Answer': '<answer>', 'Related Objects': ['shape': '<shape>', 'size': '<size>', 'color': '<color>', ...]

User:

What is the size of the object that is occluded by the truck?

GPT-40:

{'Reasoning': 'The image contains several objects: a gray sedan, a white school bus, a red tandem bike, a red road bike, a red mountain bike, a purple jet, a gray wagon, and a red cruiser. The truck is occluding a small red road bike.', 'Answer': 'Small'}

Figure 12. The prompts used for L4- Occlusion questions. In the instruction, we encourage the model to first describe all the objects in the scenes and the reasoning process of answer the question. We format the reply into a dictionary format, where the answer can be decoded and compared with the ground truth answer.

Prompts for Level 4 - 3D Pose questions

System:

You are an intelligent chatbot designed to answer questions based on an image. Your task is to analyze the images, **identify attributes of multiple objects and their facing direction in 3D space from the camera view**, and then determine the answer to the question.

INSTRUCTIONS:

Step 1: Identify the objects mentioned in the question, noting each object's shape, color, size, and direction which facing at. Objects can have a shape (e.g., 'airliner'), size ('small' or 'large'), color (e.g., 'blue'), and direction ('left', 'right', 'front', 'back') in camera view. The possible colors are: gray, blue, purple, brown, green, cyan, red, and yellow.

Step 2: Use the identified details to form a brief reasoning process for answering the question. Based on this reasoning, provide your answer in one of the following formats: (1) a phrase chosen from the available options: Small, Large, Airliner, Dirtbike, Road bike, Tandem bike, Suv, Wagon, Scooter, Mountain bike, Minivan, Sedan, School bus, Fighter, Chopper, Double bus, Truck, Articulated bus, Cruiser, Jet, Utility bike, Regular bus, Biplane, Gray, Blue, Purple, Brown, Green, Cyan, Red, Yellow, Left, Right, Front, Back, (2) an integer between 0-10 for questions like 'How many' or 'What is the number of,' or (3) 'Yes' or 'No' for questions asking 'Is there.'

Output your response using this JSON format: 'Reasoning': '<reasoning>', 'Answer': '<answer>', 'Related Objects': ['shape': '<shape>', 'size': '<size>', 'color': '<color>', 'direction': '<direction>', ...]

User:

Which direction the suv is facing?

GPT-40:

'Reasoning': 'The image contains several objects: a small gray SUV, a small gray wagon, a small white school bus, a small red sedan, a small purple jet, a small red tandem bike, a small red road bike, a small red mountain bike, and a small red utility bike. The SUV is positioned towards the left side of the image.', 'Answer': 'Left'

Figure 13. The prompts used for Level 4 - 3D Pose questions. In the instruction, we encourage the model to first describe all the objects in the scenes and the reasoning process of answer the question. We format the reply into a dictionary format, where the answer can be decoded and compared with the ground truth answer.

Prompts for Level 5 - Collision questions

System:

You are an intelligent chatbot designed to answer questions based on an image. Your task is to analyze the images, identify attributes of multiple objects and their potential collision given the assumption of moving direction in **3D space**, and then determine the answer to the question.

INSTRUCTIONS:

Step 1: Identify the objects mentioned in the question, noting each object's shape, color, size, and direction which facing at. Objects can have a shape (e.g., 'airliner'), size ('small' or 'large'), color (e.g., 'blue'), **and direction ('left', 'right', 'front', 'back') in camera view.** The possible colors are: gray, blue, purple, brown, green, cyan, red, and yellow.

Step 2: Use the identified details to form a brief reasoning process for answering the question. Based on this reasoning, provide your answer in one of the following formats: (1) a phrase chosen from the available options: Small, Large, Airliner, Dirtbike, Road bike, Tandem bike, Suv, Wagon, Scooter, Mountain bike, Minivan, Sedan, School bus, Fighter, Chopper, Double bus, Truck, Articulated bus, Cruiser, Jet, Utility bike, Regular bus, Biplane, Gray, Blue, Purple, Brown, Green, Cyan, Red, Yellow, Left, Right, Front, Back, (2) an integer between 0-10 for questions like 'How many' or 'What is the number of,' or (3) 'Yes' or 'No' for questions asking 'Is there.'

Output your response using this JSON format: 'Reasoning': '<reasoning>', 'Answer': '<answer>', 'Related Objects': ['shape': '<shape>', 'size': '<size>', 'color': '<color>', 'direction': '<direction>', ...]

User:

There is a small bicycle that is behind the aeroplane; how many tiny purple fighters are on the right side of it?

GPT-40:

{'Reasoning': 'In the image, there is a small bicycle behind the airplane. There is one small purple fighter on the right side of the small bicycle.', 'Answer': '1'}

Figure 14. The prompts used for Level 5 - Collision questions. In the instruction, we encourage the model to first describe all the objects in the scenes and the reasoning process of answer the question. We format the reply into a dictionary format, where the answer can be decoded and compared with the ground truth answer.

Prompts for Level 5 - 6D Spatial questions

System:

You are an intelligent chatbot designed to answer questions based on an image. Your task is to analyze the images, **identify attributes of multiple objects and their spatial relationship from objects' perspective in 3D space**, and then determine the answer to the question.

INSTRUCTIONS:

Step 1: Identify the objects mentioned in the question, noting each object's shape, color, size, and direction which facing at. Objects can have a shape (e.g., 'airliner'), size ('small' or 'large'), color (e.g., 'blue'), **and direction ('left', 'right', 'front', 'back') in camera view.** The possible colors are: gray, blue, purple, brown, green, cyan, red, and yellow.

Step 2: Use the identified details to form a brief reasoning process for answering the question. Based on this reasoning, provide your answer in one of the following formats: (1) a phrase chosen from the available options: Small, Large, Airliner, Dirtbike, Road bike, Tandem bike, Suv, Wagon, Scooter, Mountain bike, Minivan, Sedan, School bus, Fighter, Chopper, Double bus, Truck, Articulated bus, Cruiser, Jet, Utility bike, Regular bus, Biplane, Gray, Blue, Purple, Brown, Green, Cyan, Red, Yellow, Left, Right, Front, Back, (2) an integer between 0-10 for questions like 'How many' or 'What is the number of,' or (3) 'Yes' or 'No' for questions asking 'Is there.'

Output your response using this JSON format: 'Reasoning': '<reasoning>', 'Answer': '<answer>', 'Related Objects': ['shape': '<shape>', 'size': '<size>', 'color': '<color>', 'direction': '<direction>', ...]

User:

There is a aeroplane; how many tiny brown bicycles will it collide if it moving forward?

GPT-40:

{'Reasoning': 'The image contains a small purple jet, a small gray sedan, a small gray wagon, a small white school bus, a small red tandem bike, a small red road bike, a small brown road bike, a small brown mountain bike, and a small red sedan. The small purple jet is positioned at the back of the image. If it moves forward, it will collide with the small brown road bike and the small brown mountain bike.', 'Answer': '2'}

Figure 15. The prompts used for Level 5 - 6D Spatial questions. In the instruction, we encourage the model to first describe all the objects in the scenes and the reasoning process of answer the question. We format the reply into a dictionary format, where the answer can be decoded and compared with the ground truth answer.