A. 3DSRBench Data Card

We employ ten annotators to annotate a total of 2,772 unique visual question-answer pairs across 12 question types. We follow the annotation principles as discussed in Sec. 3.1 and adopt an two-stage pipeline to ensure various criteria are met. Specifically after annotations are collected in the first stage, we review the quality of the collected data and reject samples with low quality or ones that lead to imbalanced data distribution. Additional new annotations are collected if necessary. Furthermore, we collect human responses for all visual question-answer pairs and disregard samples that don't reach consensus by human annotators.

Dataset statistics. We annotate a total of 2,100 questions on natural images from MS-COCO [36] and 674 questions on synthetic images rendered from HSSD dataset [31]. With testing data augmentation and FlipEval Sec. 4.1, we obtain a total of 5,250 questions in 3DSRBench-real and 1,692 questions in 3DSRBench-synthetic. These questions are further evaluated with CircularEval [40].

Ethics. We follow the ethics guidelines and obtained Institutional Review Board (IRB) approvals prior to the start of our work. We described potential risks to the annotators, such as being exposed to inappropriate images from the MS-COCO dataset, and explained the purpose of the study and how the collected data will be used. All annotators are paid by a fair amount as required at our institution.

License. Our dataset is released under the CC-BY-4.0 license.

B. Baseline Models

Proprietary LMMs. To analyze the 3D spatial reasoning capabilities of state-of-the-art LMMs, we explore a variety of proprietary LMMs, *e.g.*, QWen-VL [7], Claude [4], Gemini [22], and GPT-40 [28]. For Gemini-2.0-Flash, we further evaluate two variants: (1) for Gemini-2.0-Flash-bbox we first ask the model to detect the 3D bounding box and then prompt the model to answer the question based on the 3D bounding boxes, and (2) for Gemini-2.0-Flash-think we evaluate the thinking model where the model performs deep thinking prior to answering the question.

LLaVA-v1.5 [37]. LLaVA-v1.5-7B is a strong opensourced LMM baseline built on a Vicuna-v1.5 LLM and a CLIP-ViT-L vision encoder. It extends the visual instruction tuning framework [38] with an MLP connector and a scaled up image resolution. **Cambrian-1** [54]. Cambrian-1 is a strong LMM with vision-centric designs. It features an advanced connector design, spatial vision aggregator (SVA), and high-quality visual-instruction tuning data.

LMMs with various vision encoder designs. We further experiment on a family of LMMs, extending the LLaVA-v1.5 baseline with various visual encoders design. We study the impact of the 3D awareness of visual encoders on the final 3D spatial reasoning capabilities. We adopt the same LLM and training strategy, exploring: (i) different mixed encoders: involving a second visual encoder besides CLIP, e.g., DINOv2 [47], MAE [24], and SAM [33]; and (ii) different visual connectors: a standard MLP connector and spatial vision aggregator (SVA) [54].

C. Qualitative Examples of 3DSRBench

We present two example questions for each of the 12 question types in Fig. 6 (height and location questions), Fig. 7 (orientation questions), and Fig. 8 (multi-object reasoning questions).

D. Qualitative Examples of Common and Uncommon Viewpoints

We present some qualitative examples of 3DSRBench-synthetic with multi-view images rendering the same scene and objects but from common (left) and uncommon (right) viewpoints in Fig. 9.

E. Failure Cases

We present some failures cases of GPT-40 [28] in Fig. 10 and of Gemini 2.0 Flash thinking [22] in Fig. 11.

Height



- Q: Consider the real-world 3D locations of the objects. Which object is higher in 3D world space, the cyclist in orange suit or the yellow board?
- A: The yellow board.



- Q: Consider the real-world 3D locations of the objects. Which object is higher in 3D world space, the bicycle traffic light or the blue forward sign?
- A: The blue forward sign.

Location



- Q: Consider the real-world 3D locations of the objects. Is the man with a suitcase next to or far from the fire hydrant?
- A: The man is far from the fire hydrant.



- Q: Consider the real-world 3D locations of the objects. Is the couple next to or far from the ocean waves?
- A: The couple is far from the ocean waves.



- Q: Consider the real-world 3D locations of the objects. Are the street lights directly above the train?
- A: No.



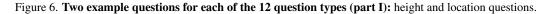
- Q: Consider the real-world 3D locations of the objects. Is the blue sign directly above the pedestrian crossing?
- A: Yes.



- Q: Consider the real-world 3D location of the objects. Which is closer to the camera, the man in grey jacket or the man in blue suit?
- A: The man in grey jacket is closer to the camera.



- Q: Consider the real-world 3D location of the objects. Which is closer to the camera, the white pickup truck or the cargo trailer?
- A: The white pickup truck is closer to the camera.



Orientation



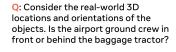
- Q: Consider the real-world 3D locations and orientations of the objects. Which side of the octagonal sign is facing the camera?
- A: The back of the octagonal sign is facing the camera.



- Q: Consider the real-world 3D locations and orientations of the objects. Which side of the bench is facing the camera?
- A: The back of the bench is facing the camera.



- Q: Consider the real-world 3D locations and orientations of the objects. If I stand at the man's location facing where it is facing, is the bottled water on the left or right side of me?
- A: The bottled water is on the right side of the man.



A: The airport ground crew is in front of the baggage tractor.





- Q: Consider the real-world 3D locations and orientations of the objects. If I stand at the taxi's location facing where it is facing, is the steel gate on the left or right side of me?
- A: The steel gate is on the left side of the taxi.
- Q: Consider the real-world 3D locations and orientations of the objects. Is the motorcycle handle in front or behind the cat?
- A: The motorcycle handle is behind the cat.



Figure 7. Two example questions for each of the 12 question types (part II): orientation questions.

Multi-Object



- Q: Consider the real-world 3D locations and orientations of the objects. Which object is closer to the fire hydrant, the bus or the potted plants?
- A: The potted plants are closer to the fire hydrant.
- Q: Consider the real-world 3D locations and orientations of the objects. Is the air conditioner facing towards the potted plants or the sofa?
- A: The air conditioner is facing towards the potted plants.



- Q: Consider the real-world 3D locations and orientations of the objects. Which side of the horse is facing towards the red sign?
- A: The right side of the horse is facing towards the red sign.



- Q: Consider the real-world 3D locations and orientations of the objects. Is the grey truck parallel or perpendicular with the stop sign.
- A: They are perpendicular to each other.



- Q: Consider the real-world 3D locations and orientations of the objects. Is the yellow cow facing the same direction as the red car?
- A: Yes, they are facing different directions.



- Q: Consider the real-world 3D locations and orientations of the objects. Which object is closer to the coffee table, the TV or the surfboard?
- A: The TV is closer to the coffee table.



- Q: Consider the real-world 3D locations and orientations of the objects. Is the dog with orange shoes facing towards the man or the dog without shoes?
- A: The dog with orange shoes is facing towards the dog without shoes.



- Q: Consider the real-world 3D locations and orientations of the objects. Which side of the dog is facing towards the speakers?
- A: The left side of the dog is facing towards the speakers.



- Q: Consider the real-world 3D locations and orientations of the objects. Is the blue bridge parallel or perpendicular to the train with orange stripes?
- A: They are perpendicular to each other.



- Q: Consider the real-world 3D locations and orientations of the objects. Is the tram facing the same direction as the arrow on the black sign?
- A: Yes, they are facing the same direction.

Figure 8. Two example questions for each of the 12 question types (part III): multi-object reasoning questions.

Common Viewpoints Uncommon Viewpoints

 $Figure \ 9. \ \textbf{Qualitative examples of multi-view images rendering the same scene and objects but from common (left) and uncommon (right) viewpoints.}$



Question: Which object closer to the camera, the fruits on the ground or the white pickup truck?

Ground-truth: The white pickup truck is closer.

GPT-40: The fruits on the ground are closer to the camera than the white pickup truck. You can tell this by their larger size and more detailed appearance in the foreground, while the pickup truck is farther away and appears smaller in comparison.



Question: Are the white sign and the stop sign parallel or perpendicular to each other?

Ground-truth: The fireplace is behind you.

GPT-40: In the image, the white sign saying "DO NOT BLOCK INTERSECTION" and the red stop sign are perpendicular to each other. The white sign is oriented vertically along the pole, while the stop sign faces outward toward approaching traffic, creating a right angle between their surfaces.



Question: If I stand at the location of the red bus facing where it is facing, is the red sign on my left side or my right side?

Ground-truth: The red sign would be on your right side.

GPT-40: If you stand at the location of the red bus facing the same direction as the bus, the red sign (the "no entry" sign visible on the left side of the image) will be on your left.



Question: If I stand at the location of the woman facing where she is facing, is the wood barrel on my left side or my right side?

Ground-truth: The fireplace is behind you.

GPT-40: If you stand at the location of the woman in the image, facing in the same direction she is, the wood barrel would be on your right side.

Figure 10. **Failure cases of GPT-4o [28] on our 3DSRBench.** We find that GPT-4o cannot perform rigorous 3D spatial reasoning. In the top figure, GPT-4o resort to visual cues for 3D spatial reasoning. In the bottom two examples, GPT-4o understands the question, attempts to perform 3D spatial reasoning from the scene, and fails to derive the correct answer.



Question: Consider the real-world 3D locations and orientations of the objects. Which side of the horse is facing towards the red sign? Answer with front, left, back, or right.

Thinking: Here's a thinking process to determine which side of the horse is facing the red sign:

- 1. Identify the key objects: The question focuses on the "horse" and the "red sign".
- Locate the red sign: The red sign is circular with a white horizontal bar, indicating a "Do Not Enter" or similar prohibition. It's mounted on a pole in the middle ground, slightly to the left of the horse.
- 3. Locate the horse: The horse is the large animal pulling the carriage in the foreground.
- 4. <u>Determine the horse's direction of movement:</u> The horse is clearly moving from left to right across the frame. This can be inferred from its posture, the way its legs are extended, and the slight blur suggesting motion.
- 5. Visualize the horse's orientation relative to the red sign:
 - The red sign is on the left side of the horse's path.
 - As the horse moves to the right, its left side will be closer to the red sign.
- 6. Consider the horse's body: Imagine standing where the red sign is. Which side of the horse would be closest to you?
- 7. <u>Confirm with visual cues</u>: While the horse is in motion, we can see the left side of its body, including its left legs and flank, are oriented generally towards the area where the red sign is located.
- 8. Eliminate other possibilities:
 - Front: The horse's front is pointing towards the right.
 - Back: The horse's back is trailing behind it, away from the red sign.
 - Right: The horse's right side is facing away from the red sign.
- 9. Formulate the answer: State clearly that the left side of the horse is facing the red sign.

Answer: Here's my reasoning to determine which side of the horse is facing the red sign:

- 1. Identify the objects: We need to locate the horse and the red sign.
- 2. Locate the red sign: The red sign is circular with a white horizontal bar in the middle, indicating a "Do Not Enter" or similar sign. It's on a pole on the left side of the image.
- Locate the horse: The horse is the large animal pulling the carriage in the foreground, moving from left to right across the image.
- 4. Determine the horse's direction of movement: The horse is moving from the left side of the frame towards the right. We can see this by its posture and the direction its legs are moving.
 5. Visualize the horse's position relative to the red sign: As the horse moves to the right, the red sign is located on its left
- side.
 6. Confirm with visual cues: We can see the left side of the horse's body, including its left legs and flank, are facing generally towards the location of the red sign.

Conclusion: The left side of the horse is facing towards the red sign.

Ground truth: Right.

Figure 11. **Failure cases of Gemini 2.0 Flash thinking [22] on our 3DSRBench.** In this example Gemini 2.0 Flash thinking again breaks down the 3D spatial reasoning question into small and tractable steps. It also derives visual cues that can help verify the answer. However, the model still fails to perform 3D spatial reasoning over multiple 3D information, *e.g.*, depth and orientation, and predicts a wrong answer.