
SD-VLM: Spatial Measuring and Understanding with Depth-Encoded Vision-Language Models

Supplementary Materials

Anonymous Author(s)

1 A Proof of Mapping from 2D to 3D with Distance Constraints

2 In this section, we prove that given an image with its depth map and enough annotated physical
3 lengths, the mapping from the image to its 3D structure can be established.

4 **Mapping from a 2D image to 3D points.** Given an image \mathcal{I} , there exists a mapping \mathcal{F} transforming
5 the image \mathcal{I} to 3D points \mathcal{P} , i.e. $\mathcal{P} = \mathcal{F}(\mathcal{I})$. For any homogeneous pixel coordinates $\mathbf{p} = [u, v, 1]^T$
6 on the image, the corresponding metric depth value d and the camera intrinsics are required for
7 mapping \mathbf{p} to the 3D point $\mathbf{P} = [X, Y, Z]^T$:

$$\mathbf{P} = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = d \cdot \mathbf{K}^{-1} \mathbf{p} = \begin{bmatrix} (u - c_x) \frac{d}{f_x} \\ (v - c_y) \frac{d}{f_y} \\ d \end{bmatrix}. \quad (1)$$

8 \mathbf{K} is the camera intrinsic matrix with four unknown parameters,

$$\mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \quad (2)$$

9 **Constraints Based on Metric Distances.** Camera intrinsics are necessary for mapping depth map
10 to 3D structure. 3D point coordinates in a camera coordinate system are hard to obtain in daily
11 scenarios, while measuring metric distances between two points is feasible.

12 Suppose two pixels \mathbf{p}_1 and \mathbf{p}_2 are the endpoints of a line segment with physical length L in the image
13 \mathcal{I} . The corresponding depth values are d_1 and d_2 and 3D points from the mapping are \mathbf{P}_1 and \mathbf{P}_2 .
14 Hence, the physical length is calculated by

$$L^2 = \|\mathbf{P}_1 - \mathbf{P}_2\|_2^2. \quad (3)$$

15 Explicitly, the constraint based on the metric distance is

$$L^2 = \left(\frac{u_1 - c_x}{f_x} d_1 - \frac{u_2 - c_x}{f_x} d_2 \right)^2 + \left(\frac{v_1 - c_y}{f_y} d_1 - \frac{v_2 - c_y}{f_y} d_2 \right)^2 + (d_1 - d_2)^2. \quad (4)$$

16 This is one nonlinear equation for four unknowns f_x, f_y, c_x, c_y . Suppose there are N line segments
17 labeled with physical lengths in an image. We define

$$E_i(f_x, f_y, c_x, c_y) = \|\mathbf{P}_{i1} - \mathbf{P}_{i2}\|_2^2 - L_i^2, \quad i = 1, \dots, N. \quad (5)$$

18 The residual vector is

$$\mathbf{r} = [E_1, \dots, E_N]^T. \quad (6)$$

19 Hence, the intrinsic parameters are estimated by optimization algorithms with the objective:

$$\min_{f_x, f_y, c_x, c_y} \|\mathbf{r}\|^2. \quad (7)$$

20 Solving above optimization problem needs at least four segments ($N \geq 4$) with ground truth length.
21 Actually, the depth values are sometime noisy and the ground-truth depth map is difficult to acquire
22 in daily scenarios. Besides, if the given depth map is relative, i.e. $d = a \cdot d_{rel} + b$, there exists two
23 more parameters. Empirically, abundant constraints would produce a more robust estimation, which
24 improves robustness to noise and improve stability.

25 Theoretically, given an image with its depth map, we can fully mapping pixels to corresponding 3D
26 point cloud when enough annotated physical lengths are provided. We believe that providing enough
27 physical labeled lengths in images would facilitate the spatial understanding of images.

28 B Details for Data Generation

29 In this section, we provide more details for the data generation procedure.

30 B.1 Prompt Details

31 **Prompts for semantic disambiguation.** We crop instances from an image, which are fed to Qwen-
32 2.5-VL with the prompt as below:

Describe the object class in the image and directly return a term. For example, the red car,
the wooden table, the man in white.

Output:

34 **Prompts for CoT data generation.** To elicit reasoning paths with reference objects, we randomly
35 select an object as the reference object, combining its spatial annotations and the image as inputs to
36 Qwen2.5-VL with the prompt as below:

Please help me rephrase the following VQA (Visual Question Answering) pairs to improve
their rationale. I will give you an image which shows an indoor environment and contains
various objects. Based on the image, I will also give you a question and answer, the question
containing a reference object. You need to propose a robust step-by-step plan to answer the
question by using the reference scales and the information from the image.

For example:

Q: The height of the chair is 0.7 m, can you measure the height of the wooden table?

A: Since the height of the chair is 0.7 m, I think the height of the wooden table is 1.4 m

Example Output:

To determine the height of the table. I should think it step by step carefully. Firstly, the
chair height can be used as reference, which is known as 0.7 m. The wooden table appears
to be about double the height of the counter. So, the height of the wooden table is 1.4 m.

Please process the following VQA pairs in the same way:

Q: [Q]

A: [A].

Output:

38 **Prompts for CoT data quality assessment.** We employ a large language model, DeepSeek-V3, to
39 evaluate and filter the CoT pairs based on their factual accuracy and logical coherence. The relevant
40 prompts are provided below:

Please help me evaluate the factual consistency and logical coherence between the original VQA pairs and the generated ones. The goal is to ensure that the generated answers align with the original facts and maintain logical reasoning.

Task:

Compare the original VQA pair with the generated one.

Check for factual consistency: Ensure that the generated answer does not contradict the original facts.

Check for logical coherence: Ensure that the generated answer and its reasoning (if provided) are logically sound and aligned with the original context.

Finally give a score between 0 and 10, where 0 indicates a poor match and 10 indicates a perfect match.

Input:

Original VQA Pair:

Q: [Original Question]

A: [Original Answer]

Generated VQA Pair:

Q: [Generated Question]

A: [Generated Answer]

The output should follow the format:

Factual Consistency: [Yes/No]

Logical Coherence: [Yes/No]

Score: [Score]. An example of output is

Factual Consistency: Yes,

Logical Coherence: Yes,

Score: 10.

Now return your output:

41

42 B.2 Statistics of MSMU

43 We categorize the spatial tasks in MSMU into 8 types, the distribution of which is illustrated in Figure
44 1 (left). The QA distribution of MSMU-Bench is also shown in Figure 1 which provides a detailed
45 breakdown of these eight categories.



Figure 1: The left shows the QA distribution of the MSMU dataset. The right shows the QA distribution of MSMU-Bench and the specific numbers of each category.

46 B.3 QA Templates

47 We provide templates that are used to construct the spatial tasks. Since MSMU consists of eight
48 spatial tasks, the templates can also be grouped into eight types, which are demonstrated as below:

49 Scale Estimation

```
1 # Scale Estimation
2 size_template_questions = [ "What is the size of [A]?", "How big is [A]?", "Can you provide the size measurement of [A] ?", ]
3 size_template_answers = [ "The size of [A] is [Length] x [Width] x [Height]. ", "[A] is with the length of [Length], width of [Width], and height of [Height]."]
4
5 height_template_questions = [ "What is the height of [A]?", "How tall is [A]?", "Can you measure the height of [A] ?", ]
50 height_template_answers = [ "The height of [A] is [Height].", "[A] is with the height of [Height].", "[A] measures [Height] in height." ]
6
7 width_template_questions = [ "What is the width of [A]?", "Determine the width of [A].", "Can you measure the width of [A] ?", ]
8 width_template_answers = [ "The width of [A] is [Width].", "[A] is with the width of [Width].", "[A] measures [Width] in width." ]
9
```

51 Counting

```
1 # Counting
2 count_template_questions = [ "How many [A]s are there in the image ?", "what's the total number of [A]s in the image?" ]
52 count_template_answers = [ "There are [X] [A]s.", "There are [X] [A]s in the image.", "[X]."]
3
```

53 Grounding

```
1 # Grounding
2 position1_template_questions = [ "What object is located at ([x],[y])?", "What can you find at ([x],[y])?", "What object does the position ([x],[y]) belong to?", ]
54 position1_template_answers = [ "It is [A].", "That is [A].", '[A].']
3 position2_template_questions = [ "What is the coordinate of [A] ?", ]
4 position2_template_answers = [ "([x],[y]).", "It is located at ([x],[y]) in the image." ]
5
```

55 Existence

```
1 # Existence
2 zero_template_questions=[ "What is the size of [A]?", "How big is [A]?", "Can you provide the size measurement of [A] ?", "What is the height of [A]?", "How tall is [A]?", "Where is [A]?",
56 "How many [A]s are there in the image ?", "what's the total number of [A]s in the image?"]
3 zero_template_answers=['There is no [A] in the image', 'Can not find [A].', 'No [A] in the scene.']
4
```

57 Absolute Distance

```
1 # Absolute Distance
2 distance_template_questions = [ "What is the distance between [A] and [B]?", "How far away is [A] from [B]?", "Can you provide the distance measurement between [A] and [B]?", ]
58 distance_template_answers = [ "[A] and [B] are [X] apart.", "A distance of [X] exists between [A] and [B].", "[A] and [B] are [X] apart from each other.", "The distance is [X]."]
3
```

59 Relative Position

```
1 # Relative Position
2 left_template_questions = [ "Is [A] to the left/right of [B] from the
    viewer's perspective?", "Does [A] appears on the left/right side
    of [B]?", "Can you confirm if [A] is positioned to the left/right
    of [B]?", ]
3 left_template_answers = ["Yes, [A] is to the left/right of [B].", "
    Indeed, [A] is positioned on the left/right side of [B]."]
60 4
5 closer_template_questions=["From the viewer's perspective, what is
    closer, [A] or [B] ?"]
6 closer_template_answers=["[X] is more closer."]
7
8 stands_template_questions=["Which stands higher/lower in the image, [A
    ] or [B] ?"]
9 stands_template_answers= ["[X] stands higher/lower."]
```

61 Scale Comparison

```
1 # Scale Comparison
2 taller_template_questions=["Between [A] and [B], which one is taller/
    lower?", "Which one is taller/lower, [A] or [B]? "]
3 taller_template_answers=["The height of [A] is [Height A] and [B] is [
    Height B], so [X] is taller/lower."]
4
5 tallest_template_questions = ["What is tallest/lowest among [A], [B],
    and [C]?"]
62 6 tallest_template_answers = ["The height of [A] is [Height A], height
    of [B] is [Height B], and height of [C] is [Height C], so the
    tallest is [X]."]
7
8 larger_template_questions=["Between [A] and [B], which one is larger/
    smaller?", "Which one is larger/smaller, [A] or [B]? "]
9 larger_template_answers=["The size of [A] is [Length A] x [Width A] x
    [Height A] and [B] is [Length B] x [Width B] x [Height B], so [X]
    is larger/smaller."]
```

63 Reference Object Estimation

```
1 # two objects
2 refer1_template_questions = [ "The height of [A] is [Height A], can
    you measure the height of [B]?"]
3 refer1_template_answers = ["Since the height of [A] is [Height A], i
    think [B] is [Height B] in height.", ]
4
5 refer2_template_questions = ["The width of [A] is [Width A], can you
    measure the width of [B]?"]
64 refer2_template_answers = ["Since the width of [A] is [Width A], i
    think the width of [B] is [Width B]"]
7
8 refer3_template_questions = ["The height of [A] is [Height A], can you
    measure the size of [B]?"]
9 refer3_template_answers = ["Since the height of [A] is [Height A], i
    think the size of [B] is [Length B] x [Width B] x [Height B]."]
```

```

1 # three objects
2 refer4_three_template_questions=["The height of [A] is [Height A],
   what is the height of [B] and [C] ?"]
3 refer4_three_template_answers=["Since the height of [A] is [Height A],
   i think the height of [B] is [Height B] and the height of [C] is
   [Height C]."]
65
4
5 refer5_three_template_questions=["The distance between [A] and [B] is
   [dis A2B], what is the distance between [B] and [C] ?"]
6 refer5_three_template_answers=["Since the distance between [A] and [B]
   is [dis A2B], i think the distance between [B] and [C] is [dis
   B2C]."]

```

C Ablation Study on Estimated Depth

Training Setting	Inference Setting	MSMU-Bench
w/ GT depth	w/ GT depth	57.71
w/ GT depth	w/ estimated depth	56.31
w/ estimated depth	w/ GT depth	54.17
w/ estimated depth	w/ estimated depth	55.35
w/o any depth	w/o any depth	46.73

Table 1: Ablation study on the sources of depth maps.

We have conducted a further investigation about the sources of depth maps. From the data presented in the Table 1, it is evident that the use of ground-truth depth maps during both training and inference phases leads to the best performance (57.71%) on the MSMU-Bench dataset. This suggests that the accuracy of depth information is crucial for the model’s ability to process and interpret spatial data effectively. If the ground-truth depth is not provided, the overall success rate of the model with estimated depth maps is still competitive. It is noteworthy that a significant disadvantage can be observed when the model is not equipped with any depth information, revealing the importance of incorporating depth priors into the VLM framework.

D Ablation Study on Normalization

To bridge the gap between different sources of depth maps, we conduct normalization in the depth map before depth positional encoding, which can be formulated as:

$$depth_{norm} = \frac{depth - depth_{min}}{depth_{max} - depth_{min}} * \alpha, \quad (8)$$

where α represents the normalization coefficient, which restricts the maximum value of the depth map.

α	MSMU-Bench
50	49.50
100	56.31
200	53.67
500	52.98

Table 2: Ablation study on the normalization coefficient.

As shown in Table 2, the highest success rate in MSMU-Bench is achieved when α is 100.

81 E Evaluation Details

82 E.1 GPT-4 Evaluation for MSMU-Bench

83 We resort to LLMs (i.e. GPT-4-Turbo) to evaluate the results. For quantitative queries, GPT-4 extracts
84 numerical values from the responses, and we calculate the success rate using a predefined threshold.
85 The prompt used to extract numerical values is shown as below:

You should help me to evaluate the response given the question and the correct answer. You need to convert the measurement of the correct answer and response to meters. The conversion factors are as follows: 1 inch = 0.0254 meters. 1 foot = 0.3048 meters. 1 centimeter (cm) = 0.01 meters. You should output two floats in meters, one for the answer, and one for the response. If the answer or response contains more than one number for prediction, you should output the List that contains the numbers. The output should be in JSON format.

Example 1:

Question: How tall is the long brown table opposite the crossed table?

Answer: The height of the long brown table opposite the crossed table is 1.02 m.

Response: It is 2.17 meters wide.

“answer_in_meters”: 1.02, “response_in_meters”: 2.17

Example 2:

Question: what’s the total number of chairs in the image?

Answer: 2.

Response: There are 2 chairs.

“answer_in_meters”: 2, “response_in_meters”: 2

Example 3:

Question: What is the size of the dark pillow?

Answer: The dark pillow is with the size of 0.8 m x 0.63 m x 0.55 m

Response: It is 35.9 inches wide.

“answer_in_meters”: [0.78,0.63,0.55], “response_in_meters”: 0.91

Example 4:

Question: The height of the bed is 0.81 m, what is the height of the table and nightstand?

Answer: Since the height of the bed is 0.81 m, i think the height of the table is 1.02 meters and the height of the nightstand is 0.93 meters.

Response: Since the height of the bed is 0.81 m, i think the height of the table is 1.36 meters and the height of the nightstand is 0.77 meters.

“answer_in_meters”: [1.02,0.93], “response_in_meters”:[1.36,0.77]

Your Turn:

Question: [Question]

Answer: [Answer]

Response: [Pred]

87 For qualitative questions, GPT-4 scores the model’s answers between 0 and 1. The prompt is shown
88 as below:

You should help me to evaluate the response given the question and the correct answer. To mark a response, you should output a single integer between 0 and 1. 1 means that the response perfectly matches the answer. 0 means that the response is completely different from the answer. The output should be in JSON format.

Example 1:
Question: Is the blue bed to the left of the curtain from the viewer’s perspective?
Answer: Indeed, the bed is to the left of the curtain.
Response: Yes, the blue bed is positioned on the left side of the curtain.
“your_mark”: 1

Example 2:
Question: Between the wooden table and the black chair, which on is taller?
89 Answer: The wooden table is taller.
Response: The chair.
“your_mark”: 0

Example 3:
Question: What is the tallest among the table, the chair, and the curtain?
Answer: The tallest is the curtain.
Response: The curtain.
“your_mark”: 1

Your Turn:
Question: [Question]
Answer: [Answer]
Response: [Response]

90 **E.2 Q-Spatial**

91 Following the official setting, the evaluation threshold for Q-Spatial is 2.0. And the system prompt
92 used is shown below:

93 You will be provided with a question and a 2D image.
The question involves measuring the precise distance in 3D space through a 2D image.
You will answer the question by providing a numerical answer.

For example:
Question: What is the distance between the two chairs?
Answer: The minimum distance between the two speckled pattern stool chairs is 1 meter.

94 **E.3 SpatialRGPT**

95 Following the official setting, the evaluation threshold for SpatialRGPT is 1.25. The prompts used to
96 evaluate the qualitative and quantitative questions are the same as those used in MSMU-Bench.

97 **F Detailed Results on SpatialRGPT-Bench**

98 More detailed results are shown in Table 3 and Table 4. Our SD-VLM shows the best performance
99 on quantitative tasks such as Height, Vertical Distance, Horizontal Distance, and Direct Distance and
100 qualitative tasks such as Big/Small, Behind/Front, Left/Right, and Tall/Short.

Model	SpatialRGPT-Bench				
	Height	Width	Vertical Distance	Horizontal Distance	Direct Distance
GPT-4o	7.8 / 0.76	9.0 / 0.67	15.1 / 0.61	18.0 / 0.65	14.9 / 0.64
Gemini-2	42.2 / 1.63	26.2 / 0.51	12.3 / 0.68	25.0 / 3.89	9.4 / 4.75
Qwen2.5-VL-72B	31.8 / 1.38	23.8 / 0.57	8.5 / 0.71	8.0 / 0.84	9.4 / 0.70
InternVL-3-78B	38.8 / 2.42	23.7 / 0.70	24.5 / 1.04	17.0 / 0.72	13.4 / 0.72
LLaVA-1.5-7B	31.0 / 1.33	24.6 / 0.55	7.5 / 0.73	10.0 / 0.82	7.9 / 0.73
SpatialBot	28.4 / 2.16	20.5 / 0.72	6.6 / 0.70	8.0 / 3.39	2.4 / 0.87
SpatialRGPT	41.3 / 0.48	44.2 / 0.51	24.5 / 0.58	13.0 / 0.64	20.5 / 0.55
Ours	42.2 / 0.55	26.2 / 0.50	35.8 / 0.50	37.0 / 0.45	25.2 / 0.55

Table 3: Results on quantitative tasks in SpatialRGPT-Bench. We report the success rate and absolute relative error for SpatialRGPT-Bench.

Model	SpatialRGPT-Bench					
	Big/Small	Behind/Front	Left/Right	Tall/Short	Wide/Thin	Below/Above
GPT-4o	55.1	59.8	63.6	56.3	55.6	72.5
Gemini-2	54.1	44.6	59.1	67.7	55.6	64.2
Qwen2.5-VL-72B	60.2	54.3	67.0	66.7	56.7	63.3
InternVL-3-78B	60.2	55.4	64.8	66.7	60.0	65.8
LLaVA-1.5-7B	16.3	40.2	23.9	32.3	17.8	27.5
SpatialBot	61.2	45.7	59.1	58.3	54.4	56.6
SpatialRGPT	59.2	56.5	39.8	65.6	55.6	70.0
Ours	61.2	67.4	68.2	69.8	58.9	67.5

Table 4: Performance of various baselines on the qualitative spatial tasks in SpatialRGPT-Bench.

G Limitations

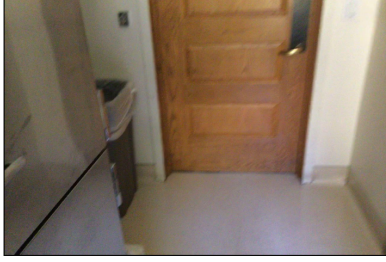
MSMU concentrates on indoor settings, featuring objects typical of domestic environments, reflecting our source datasets’ composition. It narrows the model’s applicability to social or dynamic interaction contexts. However, our model still exhibits strong adaptability, as evidenced by its solid performance on benchmarks like SpatialRGPT-Bench, which contains abundant outdoor scenes. In the future, we will explore larger base models and alternative architectures, such as Qwen-VL, to further investigate our proposed framework.

H Broader Impact

Our model enhances its role as a robust multi-modal generalist by demonstrating superior precision in spatial understanding. This capability is particularly valuable in embodied AI applications, where it aids robots in perceiving their surroundings with greater accuracy and performing precise manipulations. In addition, our model, which is based on large language models, may encounter issues with hallucination, posing significant challenges when deploying the model in real-world environments.

I More Result Comparisons on MSMU-Bench

Figure 2 and 3 illustrate more result comparisons on MSMU-Bench tasks. Our model shows a consistent advantage in spatial measuring and understanding. It is worth noting that our model is able to reason about complex spatial tasks with the chain-of-thought while other models fail to answer or return an incorrect answer, as shown in Figure 3.



Question (Existence):

Can you provide the size measurement of washing machines?

Answer:

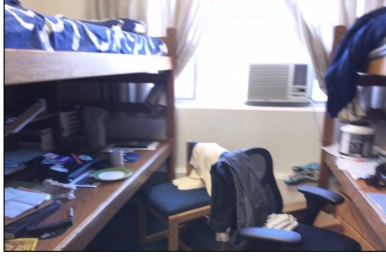
GT: *Can not find washing machines.*

GPT-4o: I'm unable to provide size measurements as there are **no** washing machines visible in the image.

LLaVA-v1.5: **N/A.**

SpatialRGPT: The size of washing machines can vary depending on the model and brand.

Ours: There is **no** washing machines in the image.



Question (Counting):

How many chairs are there in the image?

Answer:

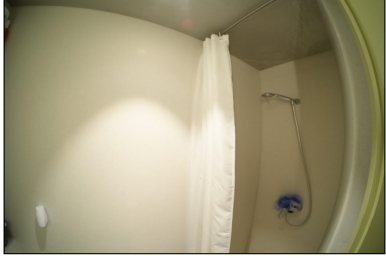
GT: 2.

GPT-4o: There are **two** chairs in the image.

LLaVA-v1.5: There are **two** chairs in the image.

SpatialRGPT: There are **3** chairs in the image.

Ours: There are **2** chairs.



Question (Grounding):

What object is located at (0.53,0.59)?

Answer:

GT: *The shower curtain.*

GPT-4o: The object located at (0.53, 0.59) is a **showerhead**.

LLaVA-v1.5: The shower curtain is located at (0.53,0.59).

SpatialRGPT: The object located at (0.53,0.59) is a **shower**.

Ours: It is the **shower curtain**.



Question (Scale Estimation):

What is the height of the office chair?

Answer:

GT: *The office chair measures 1.08 meters in height.*

GPT-4o: The height of the office chair is **0.8 meters**.

LLaVA-v1.5: The office chair is of **average height**.

SpatialRGPT: The height of the office chair is **30 inches**.

Ours: The height of the office chair is **1.06 m**.



Question (Relative Position):

Can you confirm if the coffee machine is positioned to the left of the dishwasher?

Answer:

GT: *No, the coffee machine is to the right of the dishwasher.*

GPT-4o: Yes, the coffee machine is positioned to the **left** of the dishwasher.

LLaVA-v1.5: **N/A.**

SpatialRGPT: Yes, the coffee machine is located to the **left** of the dishwasher.

Ours: No, the coffee machine is to the **right** of the dishwasher.



Question (Scale Comparison):

Between the white toilet and the shower curtain, which one is larger?

Answer:

GT: *The size of the white toilet is 0.69 m x 0.64 m x 0.71 m and the shower curtain is 41.29 cm x 30.74 cm x 183.82 cm, so the white toilet is larger.*

GPT-4o: In terms of overall size, the **shower curtain** is larger than the toilet.

LLaVA-v1.5: **N/A.**

SpatialRGPT: The **shower curtain** is larger than the toilet.

Ours: The size of the white toilet is 0.74 m x 0.49 m x 0.77 m and the shower curtain is 0.44 meters x 0.31 meters x 1.89 meters, so the **white toilet** is larger.

Figure 2: More results on MSMU-Bench.



Question (Reference Object Estimation):

The height of the right shelf is 1.65 meters, can you measure the height of the desk with laptop and water bottle?

Answer:

GT: *Since the height of the right shelf is 1.65 meters, i think the desk with laptop and water bottle is 0.71 meters in height*

GPT-4o: I'm unable to directly measure distances or heights in images. However, you can possibly estimate the desk height by comparing it to the 1.65-meter tall shelf if both are on the same plane and at a similar distance from the camera.

LLaVA-v1.5: *N/A.*

SpatialRGPT: The height of the desk with laptop and water bottle is **0.55 meters**.

Ours: To determine the height of the desk with the laptop and water bottle, I should think it through step by step carefully. Firstly, the right shelf's height is given as 1.65 meters, which can be used as a reference scale. Observing the image, the desk appears to be slightly shorter than the right shelf. By estimating the proportion, the desk seems to be around 50% of the shelf's height. Therefore, multiplying 1.65 meters by 0.50 gives an approximate height of 0.825 meters for the desk. So, the height of the desk with the laptop and water bottle is estimated to be **0.825 meters**.

Figure 3: An example of the reference object estimation question in MSMU-Bench.