

# Thinking with Blueprints: Assisting Vision–Language Models in Spatial Reasoning via Structured Object Representation

## Supplementary Material

### 1. Details of the Reinforcement Learning Stage

Group Relative Policy Optimization (GRPO) is adopted as the advantage estimator of the reinforcement learning process, which improves the stability of policy training on long-form trajectories by using group-wise normalized advantages and applying a clipped, token-level PPO-style objective.

Given a group of  $G$  trajectories  $\mathcal{O} = \tau_{i=1}^G$  conditioned on an input  $x$ , each trajectory  $\tau^{(i)}$  receives a scalar reward  $r^{(i)} = R(\tau^{(i)})$ . GRPO then forms a centered advantage  $\hat{A}^{(i)} = r^{(i)} - \bar{R}$ , where the baseline  $\bar{R} = \frac{1}{G} \sum_i r^{(i)}$  is the average reward over the group.

Let  $\tau_t^{(i)}$  denote the  $t$ -th token of trajectory  $\tau^{(i)}$ . The GRPO objective is the following clipped surrogate loss:

$$\begin{aligned} \mathcal{L}_{\text{GRPO}}(\theta) = & -\frac{1}{G} \sum_{i=1}^G \frac{1}{|\tau^{(i)}|} \sum_t \min \\ & \left[ \rho_t^{(i)} \hat{A}^{(i)}, \text{clip}(\rho_t^{(i)}, 1-\varepsilon, 1+\varepsilon) \hat{A}^{(i)} \right] \\ & + \beta \text{KL}[\pi_\theta \| \pi_{\text{ref}}]. \end{aligned} \quad (1)$$

where  $\rho_t^{(i)} = \frac{\pi_\theta(\tau_t^{(i)} | \tau_{<t}^{(i)}, x)}{\pi_{\text{old}}(\tau_t^{(i)} | \tau_{<t}^{(i)}, x)}$  is the importance sampling ratio,  $\varepsilon = 0.2$  is the clipping threshold, and  $\beta$  controls the strength of the KL regularization toward the reference policy  $\pi_{\text{ref}}$ . For the advantage estimator  $A$ , all the advantages whose rewards falls inside the standard deviation range around the mean value are zeroed out, which is formally expressed as follows.

$$\begin{aligned} \tilde{R}_i &= \frac{R(y_i) - \text{mean}(R_G)}{\text{std}(R_G)}. \\ \mathbb{1}(|R(y_i) - \text{mean}(R_G)| \geq \text{std}(R_G)), & \quad (2) \\ \hat{A}_{i,t} &= \tilde{R}_i \ (\forall t \in \{1, \dots, |y_i|\}), \\ R_G &= \{R(y_1), \dots, R(y_G)\}, \ R(y_i) \in R_G. \end{aligned}$$

Such formulation is particularly effective for stabilizing optimization in long-horizon, multimodal reasoning tasks.

### 2. Details of Data Augmentation Pipeline

Here we detail the process of our data augmentation pipeline. For SAT training set, we let the VLM (in practice gpt-4o) to decide whether to edit the image  $I$  and augment it into  $I'$  and remain the question  $Q$  unchanged, or to perturb the question  $Q$  and obtain  $Q'$  which leads to the reverse of the original answer  $a$  and remain the image  $I$  unchanged.

For augmenting the image  $I$  and augment it into  $I'$ , the process is as follows. At first, both  $Q$  and  $I$  are sent to a VLM (in practice GPT-4o) to detect the subset of the objects  $O$  in the image  $I$  that relates to the question  $Q$ . Then for each object  $o_k \in O$ , we let the VLM write an object removal prompt  $t_k$ . Finally all  $t_k$  are sent to an image editor one by one, until all the concerning objects in the image are removed.

For augmenting the question  $Q$  into  $Q'$ , the process is as follows. We combine both  $Q$  and  $I$ , along with a prompt template  $T$  which contains a set of examples of reversing the question, and send them to a VLM (in practice GPT-4o) to let the VLM decide the modified question  $Q'$ . The full prompt template about the process of editing images and questions is shown in Table 4.

Finally, all the pairs of  $(I, Q', a')$  and  $(I', Q, a')$  are sent to the VLM again so as to assure the new answer adheres to the image-question-answer triplet, and the augmented question / answer is plausible.

### 3. More Details about Experiment Settings

#### 3.1. Details of data selection in the training set.

SAT is used as training set throughout the reinforcement learning stage. In practice, to make a fair comparison with VigoRL, we conduct an even sampling with the same amount of the total QA pairs used. In particular, we randomly sample 24373 samples from SAT static, 3071 from action consequence, 2313 from action sequence, 1290 from object movement, 1233 from goal aim, and 442 from perspective.

#### 3.2. Full list of subcategories used in Blink.

For testing Blink dataset, we use 11 of the 14 categories. The full list of categories used is: Visual Correspondence, Jigsaw, Spatial Relation, Semantic Correspondence, Visual Similarity, Multi-view reasoning, Functional Correspondence, Relative Depth, Object Localization and Counting.

#### 3.3. Full prompts of proprietary models.

The prompt of GPT-4o, GPT-4o with CoT and GPT-5-Thinking are shown in Table 1 accordingly.

### 4. More qualitative examples

#### 4.1. More Visualization of Reasoning Traces.

More visualization of the whole reasoning trajectory of different methods is shown from Figure 3 to Figure 30. From

---

**Prompt template for testing proprietary models.**

```
prompt_gpt4o = "You are a careful visual QA assistant. Given an image and a multiple-choice question, answer the EXACT TEXT of the correct answer choice wrapped by <answer> and </answer>, verbatim from the options. Do not add letters (A/B), punctuation, or extra words in your answer. "  
prompt_gpt4o_cot = "You are a careful visual QA assistant. Given an image and a multiple-choice question, answer the EXACT TEXT of the correct answer choice wrapped by <answer> and </answer>, verbatim from the options. Do not add letters (A/B), punctuation, or extra words in your answer. "  
Think step by step before answering. Put your thinking process between <think> and </think> tags."  
prompt_gpt5_thinking = "You are a careful visual QA assistant. Given an image and a multiple-choice question, answer the EXACT TEXT of the correct answer choice wrapped by <answer> and </answer>, verbatim from the options. Do not add letters (A/B), punctuation, or extra words in your answer. "  

```

---

Table 1. Prompt template for testing proprietary models like gpt-4o, gpt-4o with CoT and gpt-5-thinking.

the qualitative examples we can discover that our method yields far better performance in different kinds of spatial reasoning tasks, thanks to the blueprint-based thinking patterns introduced, as well as the corresponding training strategies designed.

It is worth noting that we also discovered the inconsistency between reasoning and final answer in gpt-4o and gpt-5-thinking (Figure 19), illustrating that the inconsistency issue is agnostic among base models, thus further illustrating the importance of our design of the consistency reward.

## 4.2. About Video Spatial Reasoning.

**Method details for adapting our method to video spatial reasoning.** Here we detail about how to combine our method with a frame selector so as to expand our method to video-based spatial reasoning. Following Q-Frame, we first sample 128 frames uniformly from the input video. SigLIP is then applied to re-rank the downsampled 128 frames to select the top 4 candidate frames. Our model goes through the frames one by one. If the frame does not have the relevant information, it will say "*Questions and image do not match.*" and go to the next frame. Our model stops when it finds a frame where all the information about the question is inside the frame, or when all four candidate frames are used up but it did not find the proper answer.

### More Qualitative Results of Video Spatial Reasoning.

The qualitative illustrations about how to extend our method to video spatial reasoning is shown in Figure 1 and Figure 2. From Figure 1 we can discover that our model can quickly reach the correct answer when the correct frame is on top of the candidate list. Figure 2 also showcases that our model can exclude irrelevant frames when Q-Frames fail to rank the most relevant one on top of the list and finally navigate itself to the correct frame. Thanks to the methodological design in the adversarial images, which extends the model’s capability to distinguish irrelevant frames and push our model further to video spatial understanding.

## 4.3. Failure cases.

In this part we showcase some typical failure cases. Failure cases mostly happen at the inability to perceive all the objects in the image, lacking the world knowledge prior, as well as the ignorance of the consecutive 3D space and the camera geometry.

From Figure 31 we can observe that failures may happen when our method fails to observe all the objects in the input image. In this example, for the man in red, our model only sees the standing man which is obvious in the foreground, but it has neglected the man sitting far away in the background, looking towards the camera behind his sunglasses. This calls for the stronger grounding capabilities in VLMs. Figure 32 shows the case of the absence of real-world priors, where the model fails to realize that the sight of the luggages will be hindered by the non-transparent shells of the wagons if it entered the first wagon.

The ignorance of the consecutive 3D space and camera geometry is shown in Figure 33 and Figure 34. From Figure 33 we can discover that our model failed to perceive the camera has zoomed in, thus giving a conclusion that the lamp has moved right towards the cameras. From Figure 34 we can discover that our model becomes dizzy under the sophisticated camera rotation in 2 axis, neglecting the movement of the plant and regarding it as stationary. However, the movement of the plant can be perceived considering its distance to the desk, or by some 4D reconstruction methods.

These failure cases have shed light on future directions where we should nurture the sense of real world physical priors, real-world spatial senses, as well as the camera geometry for a better spatially-aware VLM.

## 5. More ablation studies.

### 5.1. Ablation on adding image diversity.

**Explanation of iid and ood in the experiments.** We define iid as the combination of image style and the specific task has been covered in the training set, and all other cases are referred to as ood. In this sense, as our method is trained

Method	Model Size	Reasoning Mode	SAT val ( <i>iid</i> )	SAT test ( <i>iid</i> )	Blink ( <i>ood</i> )	Robospacial ( <i>ood</i> )	VSR ( <i>iid</i> )
<i>Our Method</i>							
Ours (Trained on SAT)	7B	Thinking with Blueprint	92.7	79.7	60.7	70.2	84.8
Ours (Trained on SAT & VSR)	7B	Thinking with Blueprint	92.5	83.7	60.9	71.1	87.8
<i>Gain over adding training data diversity</i>			-0.2	+4.0	+0.2	+0.9	+3.0

Table 2. Ablation on training with both synthetic (SAT) and real images (VSR). Training on the combination of synthetic and real images helps improve generalizability of our method to real-world images. Note that here iid and ood are under the setting of dataset consists of both SAT and VSR.

Method	Model Size	Reasoning Mode	SAT val ( <i>iid</i> )	SAT test ( <i>ood</i> )	Blink ( <i>ood</i> )	Robospacial ( <i>ood</i> )	VSR ( <i>ood</i> )
<i>Proprietary Models</i>							
gpt-4o	-	No	57.2	51.5	59.2	60.1	78.7
gpt-4o	-	Naive CoT	57.7	63.3	59.0	63.6	82.2
gpt-4o	-	Thinking with blueprint	68.7	67.3	59.7	64.7	83.4

Table 3. Ablation on applying our blueprint-based thinking pattern to different models. The results of applying our methods to gpt-4o via in-context learning are reported.

on SAT training set, which is composed of synthetic indoor images, there exists a domain gap on the images between SAT training set and SAT test set, Blink, Robospacial, and VSR, where the latter are composed of real images in both indoor and outdoor scenes, say, they have OOD images. Moreover, for questions, blink contains general visual reasoning questions like semantic correspondence, and robospacial contains questions about compatibility, which are never seen throughout the training process. In this sense, Blink and robospacial also have OOD questions.

**Ablation of adding real-world images to training set.** To further examine whether the absence of real images will do harm to the generalizability to the model, we conduct an ablation study on adding real-world images at the reinforcement learning stage. We form a new training set that is composed of our sampled SAT training data, the VSR training set, and our augmented data, so that the model has familiarity with spatial reasoning on real-world images. In this ablation study, both SAT test set and VSR have become iid datasets as our model has conducted reasoning on real-world images. However, Blink and Robospacial are still OOD, as the capability in the tests set are still not experienced throughout the training process. The results are shown in Table 2. We can observe that the performance on SAT test set and VSR test set has witnessed non-trivial improvements, while the performance on SAT validation set has only negligible loss. This showcases the importance of adding the diversity of training data throughout the reinforcement learning process. The performance gain is also witnessed on Blink and Robospacial, indicating that the diversity of images can help generalization even when task

discrepancy exists. However, the improvement is not as big as SAT-test and VSR, thus calling for task diversity during the data construction stage of the training process by the side.

## 5.2. Ablation on different model architectures.

To further investigate whether our blueprint-based reasoning method can be applied to different model architectures, we conduct an ablation experiment where we put our blueprint-based thinking strategy into prompt templates and conduct in-context learning on gpt-4o. The results are shown in Table 3, where prompting GPT-4o at test time also yields performance gains, which thereby showcases the broad applicability of our blueprint-based reasoning method.

## 6. More performance comparison and analysis.

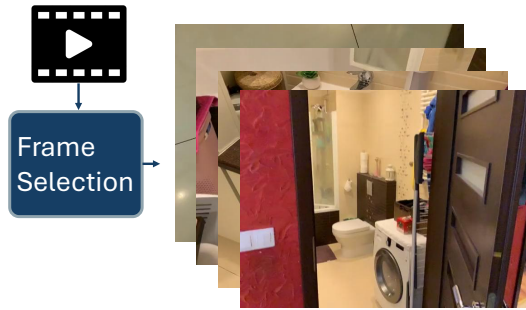
### 6.1. Additional notes to Table 1 in the main paper.

In this part we illustrate the quantitative comparison between our method and Robix and Mirage on a certain datasets, in particular, the results marked  $-^*$  in Table 1. For Mirage, as all the test sets are conducted on a private sample of size 500, where we cannot make direct comparison to without the exact samples. Also, Mirage finetunes the model on each subtask before training, such as SAT GoalAim and Blink Jigsaw, making it hard to compare with our model where it is pretrained on the all the tasks on SAT and conduct zero-shot testing on Blink without any finetuning. For Robix, as it is conducted only on blink spatial relation understanding and depth perception, we also

cannot directly compare with them about the performance. The weighted average of our method on the two subsets are 85.13%, surpassing Robix 32B and comparable with Robix 7B with far less data used during training. This showcases the effectiveness of our method by the side.

## **6.2. About failures in comparative models.**

About the failures of reasoning in competitive models, here we report the detailed statistics. For Vilasr, we have witnessed about 8.9% tool call failures, mainly composed of wrong indexing on image pixels during image operations. For gpt-5-thinking, we have witnessed 14.9% of the samples that are not successful to produce final answers. Such cases are qualitatively shown in [Figure 7](#), [Figure 10](#), [Figure 13](#), [Figure 21](#), and [Figure 29](#), where the reasoning continues without arriving at a final answer, even when it has reached correct conclusions.



**Question:** *If I am standing by the bathtub and facing the washer, is the toilet to the front or the back of the washer? Answer Choices: A. front B. back Answer with the text of the option.*

### Thinking with Blueprint

<think>I will first focus on the washer, which is positioned near the white wall to the right of the image, as it is the most distinctive object for spatial orientation. The bounding box coordinates for the washer will be located in the right region of the image.

```
{ "object": "washer", "bbox": [[240, 320], [280, 500]], "description": "A washer near the white wall." }
```

I will then focus on the toilet, as it is clearly positioned in the center left of the image. The toilet is next to the wall, along and behind the washer.

```
{ "object": "washer", "bbox": [[240, 320], [280, 500]], "description": "A washer near the white wall." },
```

```
{ "object": "toilet", "bbox": [[140, 220], [260, 360]], "description": "A toilet near the white wall." }
```

I will then focus on the bathtub, as it is partially hindered in the background.

```
{ "object": "washer", "bbox": [[240, 320], [280, 500]], "description": "A washer near the white wall." },
```

```
{ "object": "toilet", "bbox": [[140, 220], [260, 360]], "description": "A toilet near the white wall." }
```

```
{ "object": "bathtub", "bbox": [[80, 0], [100, 360]], "description": "A bathtub in the background." }
```

I have found all the objects necessary for reasoning. Now I need to do the reasoning based on the JSON and the image. The toilet and the washer are at the same side of the image, the right side. The bathtub is at the background of the image. Standing at the bathtub facing the washer means facing the camera, where the front and back are the opposite.

Given that the toilet is at the back of the washer in the image, it is at the front in the person's view. I have sufficient information from the image. I will directly make the answer.</think>

<answer>front</answer>



Figure 1. Qualitative example of applying our method to video spatial reasoning.



**Question:** *If I am standing by the bed and facing the chair, is the desk to the left or the right of the chair?*  
**Answer Choices:** A. left B. right  
**Answer with the text of the option.**

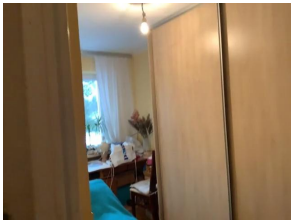
### Thinking with Blueprint



<think>I will first focus on the chair, which is positioned near the center of the image, ... .  
 [{"object": "chair", "bbox": [[20, 0], [500, 500]], "description": "A chair occupies most of the image."}]

I have found all the objects related in the image. Now I need to do the reasoning based on the JSON and the image. Based on the image, I cannot find the bed and the desk. The question and the image do not match.</think>

I do not have sufficient information from the image. I will move on to the next image.



<think>I will first focus on the chair, which is positioned near the center of the image, ... .  
 [{"object": "chair", "bbox": [[240, 380], [280, 500]], "description": "A chair at the back."}]  
 I will then focus on the bed, as it is clearly positioned in the center left of the image. ....  
 [{"object": "chair", ...}, {"object": "bed", "bbox": [[180, 300], [260, 500]], "description": "A bed behind the door."}]  
 I will then focus on the desk, as it is in the background.  
 [{"object": "chair", ...}, {"object": "bed", ...}, {"object": "desk", "bbox": [[180, 220], [24, 400]], "description": "A desk in the background."}]

I have found all the objects necessary for reasoning. Now I need to do the reasoning based on the JSON and the image. The desk is behind the chair in the image. Standing by the bed and facing the chair means the desk at behind is the left in the viewer's perspective, where the front is the right in the image.

I have sufficient information from the image. I will directly make the answer.

</think>

<answer>left</answer>



Figure 2. Qualitative example of applying our method to video spatial reasoning.

---

**Prompt Template for Augmenting Data.**

Input: Image  $I$ , Question  $Q$ , Answer  $a$ .

Output: New Image  $I'$ , New Question  $Q'$ , New Answer  $a'$ .

prompt0 = "" You are conducting an editing task over a image question pair. You are required to either remove some specific objects mentioned and inpaint the background, or to rewrite the question.

For image editing, do not hallucinate or add irrelevant objects that does not present in the background. For editing questions, make sure the question is related to the image and its correct answer.

Do not modify any irrelevant information in the image such as the image style or other objects present. ""

img\_prompt1 = "" First let's make clear what to edit. Below is a question that is not directly relevant to the editing instruction. Analyze the question text and answer the following question: What are the objects mentioned in this question or you need to observe in this image? If you think the question are not directly related to any of the objects in the image, simply answer 'No Objects'. ""

img\_prompt2 = "" From your answer to the previous question, what are the locations of the objects you have mentioned before in the image? If some objects do not appear in the image, just find the object that is the most alike in appearance. Note that some objects might have appeared multiple times. So, carefully examine this image. ""

img\_prompt3 = "" Summarize your previous answer into 30 words. One object in a line. If there is no object to be removed, simply answer Nothing. ""

text\_prompt = "" Now, based on the image, make the least modification to the question so the correct answer becomes the opposite of the original answer. Here are some templates fyi.

1. If I move to "X", will something to my left / right / will something be nearer / farther away ?

You can change into 'If I move to something, will "X" to my left or right / will something be nearer or farther away.' Now the correct answer becomes the reverse of the original answer.

2. For someone at xxx, will A be to their left or right?

You can change into 'For someone at A, will xxx be to their left or right?' Now the correct answer becomes the reverse of the original answer.

3. If I turn left / right / look straight, will I be facing away from xxx?

For left and right cases, simply reverse the direction. For the cases of look straight, change the latter part into facing towards xxx. Now the correct answer becomes the reverse of the original answer. 4. I need to go to xxx, which direction should I turn to face the object?

You can change into 'I don't want to see xxx, which direction should I turn to face away from the object?' Now the correct answer becomes the reverse of the original answer.

5. Were any of the objects in the initial frame that you can still see in the second frame moved from their original positions?

You can change into 'Were any of the objects in the second frame that you can still see in the initial frame moved from their original positions?' Now the correct answer becomes the reverse of the original answer.

6. If I rotated left and move forward, will xxx be away from the camera?

You can change 'away' into 'closer to'. Double check if the correct answer is indeed the reverse.

7. The first image is from the beginning of the video and the second image is from the end. How did the camera likely move when shooting the video?

You can change into 'The first image is from the end of the video and the second image is from the beginning. ...' Now the correct answer becomes the reverse of the original answer.

""

```
edit_decision = call_gpt4o(Image, Question, img_prompt0, img_prompt1)
```

```
if edit_decision != "No Objects":
```

```
    edit_locations = call_gpt4o(Image, Question, edit_decision, img_prompt2)
```

```
    object_summaries = call_gpt4o(Image, Question, edit_locations, img_prompt3)
```

```
    New Image = Image; New Question = Question
```

```
    for i in object_summaries:
```

```
        New Image = call_flux_kontext(New Image, i)
```

```
else:
```

```
    New Image = Image; New Question = call_gpt4o(New Image, Question, text_prompt)
```

```
New Answer = call_gpt4o.to.reverse.answer(New Image, New Question, Answer)
```

---

Table 4. Prompt template and algorithm used throughout our data augmentation process.



**Question:** How many telephone poles are to the right of the truck?

Answer Choices:

- A. 5
- B. 2
- C. 4
- D. 3

Answer with the text of the option.

### Thinking with Blueprint

<think>I will first observe the telephone pole closest to the left side of the image that is positioned near the truck. I locate this telephone pole in the left section of the image, near the edge. The bounding box coordinates for this object are `[[50, 50], [70, 200]]`.

`{{"object": "telephone pole 1", "bbox": [[50, 50], [70, 200]], "description": "A telephone pole positioned on the left side of the image near the edge, close to a truck."}}`

I then observe the telephone pole next to the truck. The bbox coords are `[[100, 100], [170, 200]]`.

`{{"object": "telephone pole 1", "bbox": [[100, 100], [170, 200]], "description": "A telephone pole at the left next to a truck."}}`

I then observe the other telephone poles. The bbox coords are `[[130, 100], [270, 200]]`.

`{{"object": "telephone pole 1", "bbox": [[130, 100], [270, 200]], "description": "Three telephone poles right to a truck."}}`

I have found all the objects necessary for reasoning. Now I need to do the reasoning based on the information I have obtained. Step 1: I first identified the telephone pole closest to the left side near the truck. This telephone pole is distinctly visible and positioned towards the left edge of the image, near a visible truck. I noted its bounding box coordinates as `[[50, 50], [70, 200]]`.

Step 2: I examined the image for additional telephone poles to the right of the truck. I observed that there are multiple telephone poles visible in the left side of the image near the road. I carefully counted these telephone poles.

Step 3: I counted the telephone poles visible to the right of the truck. I noted its bounding box coordinates as `[[130, 100], [270, 200]]`. I identified that there are exactly 3 telephone poles visible to the right of the truck, positioned along the road.

Step 4: I cross-checked my observation to ensure that I did not miss any telephone poles. The area near the road and utility lines was thoroughly analyzed, and no additional telephone poles were found beyond the 3 already identified.

Step 5: The options provided in the question are "A. 5," "B. 2," "C. 4," and "D. 3." Since I identified 3 telephone poles to the right of the truck in the image, I select the option "3."

Step 6: Finalizing the answer, I confirm that the correct choice is "D. 3," as it matches the number of telephone poles I counted in the image.</think>

<answer>3</answer>



Figure 3. Qualitative example 1. (1 of 3).



**Question:** How many telephone poles are to the right of the truck?

**Answer Choices:**

- A. 5
- B. 2
- C. 4
- D. 3

**Answer with the text of the option.**

**gpt-4o+CoT**

First, I identify the truck in the image. Then, I look to the right of the truck to count the number of telephone poles exclusively on that side, which is 2. Ensure to count only those poles to the right of the truck and not any other poles scattered elsewhere. Based on the image analysis, determine this number and match it to one of the given answer choices. The answer is 2. ❌

**gpt-5-thinking**

\*\*\*Clarifying pole counting\*\*

I need to count the telephone poles to the right of the truck. The image shows the truck on the left side, traveling down the road. So, "to the right of the truck" must mean from the viewer's perspective, but I'm thinking: there are poles visible on the left side of the road near the truck. Am I mistaken? I wonder if the question is asking for poles on the right side of the truck in the image, horizontally? \*\*Counting poles along the road\*\*

I need to zoom in and assess the sequence of poles along the road. The nearest pole is just to the right of the truck, and there seem to be more further along. If I count carefully, I see about four poles visible from the truck to the center of the image. There's a set of tall poles with crossarms and wires. I think I can make out five poles, starting from the one right next to the truck and moving further right. There's a group of four poles on the leftmost edge as well. \*\*Locating poles relative to the truck\*\*

I see there's a cluster of three slender poles at the far left edge of the image, with power lines crossing above. However, those are to the left of the truck, which isn't what I need. The question focuses on telephone poles to the right of the black pickup truck. It's in the left lane, going away. Just to the right of the truck, I can identify the first pole, then the second, third, and possibly a fourth pole as well. \*\*Counting visible poles\*\*

Upon closer inspection, I see a set of four poles on the left side of the road, extending into the distance. To the right of the truck, I can identify four distinct poles: the first is close to the truck, the second is near a sign, the third is near another sign, and the fourth is further back near the curve. There might be a fifth pole hidden near the trees, but I'll count precisely to confirm. The answer options are A. 5, B. 2, C. 4, and D. 3, and I suspect the correct count is likely 4. \*\*Marking the poles\*\*

I'm going to mark the poles I see in relation to the truck.

Pole 1 is directly to the right of the truck, near the road edge, with some equipment at the top. \n\nPole 2 is a bit further right, next to a white sign, and it's a wooden pole. \n\nPole 3 is near a second sign, and Pole 4 is further down, close to a big tree, with wires visible at the top.

I might spot a fifth pole, but I think four is the right count, so I'll return the answer as "4." ❌

**VigoRL**

<think>There is one telephone pole visible to the right of the truck at [[109, 118], [120, 194]]. There appears to be another telephone pole near the center of the image at [[133, 120], [140, 200]]. This suggests there might be more than one. There is a third telephone pole closer to the right side of the image, located at [[150, 120], [160, 200]]. This confirms that there are indeed multiple telephone poles. At this point, I can see a fourth telephone pole near the far right of the image, positioned at [[165, 120], [170, 200]]. It seems that this confirms our count so far. The fifth and final telephone pole is located towards the left side of the image, near the center, at [[125, 118], [130, 194]]. This confirms that there are five telephone poles in total.</think>

<answer>five</answer> ❌

Figure 4. Qualitative example 1. (2 of 3).



**Question:** How many telephone poles are to the right of the truck?

**Answer Choices:**

- A. 5
- B. 2
- C. 4
- D. 3

**Answer with the text of the option.**

### ViLASR

<think> To determine how many telephone poles are to the right of the truck, I should first identify the truck and the telephone poles in the image. The truck is likely the dark vehicle on the left side of the image, and the telephone poles can be seen along the side of the road. I'll use the Object Mapper tool to map these objects and see which telephone poles are to the right of the truck.

Action: Object Mapper [{"index": 1, "bbox\_2d": [78, 163, 105, 189], "label": "truck"}],

The truck is correctly identified. Now, I should identify the telephone poles and see which ones are to the right of the truck. The telephone poles are the tall objects with 橫樑 on the left side of the image. I'll use the Object Mapper tool again to map the telephone poles.

Action: Object Mapper [{"index": 2, "bbox\_2d": [2, 29, 34, 219], "label": "telephone pole"}, {"index": 2, "bbox\_2d": [60, 60, 91, 184], "label": "telephone pole"}],

The telephone poles are correctly identified, but the relative positioning to the truck needs to be verified. The truck is on the left side of the image, and the telephone poles are also on the left. The relative positioning tool can help confirm the relative position of the telephone poles to the truck.

Action: Path Tracer [{"index": 3, "start\_point\_2d": [87, 175], "end\_point\_2d": [48, 175], "label": "from truck to telephone pole"}],

The path tracer shows the relative position of the telephone pole to the truck. The path tracer shows that the telephone pole is indeed to the right of the truck. Now, I need to confirm the count of such telephone poles.</think>

Action: Answer\n<answer>B</answer>



Figure 5. Qualitative example 1. (3 of 3).



Image 1

Image 2

Image 3

**Question:** Given the first image with the lower right corner missing, can you tell which one of the second image or the third image is the missing part? Imagine which image would be more appropriate to place in the missing spot. You can also carefully observe and compare the edges of the images. **Answer Choices:** A. the second image B. the third image **Answer with the text of the option.**

### Thinking with Blueprint

<think>I will first observe the first image (Image 1), focusing on the lower-right corner where the object is missing. The bounding box coordinates for this empty space will be  $[[400, 100], [500, 200]]$ , and I will name this object "Missing portion." The description of this object is: "A rectangular area in the bottom-right corner of Image 1 that is entirely absent, leaving a black square in its place."

`[{"object": "Missing portion", "bbox":  $[[400, 100], [500, 200]]$ , "description": "A rectangular area in the bottom-right corner of Image 1 that is entirely absent, leaving a black square in its place."}]`

I have found all the objects necessary for reasoning. Now I need to do the reasoning based on the information I have obtained. Step 1: I observe Image 1, which shows a partially damaged building with a black square in the bottom-right corner indicating a missing portion. This missing portion occupies a rectangular area in Image 1.

Step 2: I analyze Image 2, which shows a scene with several people walking in front of buildings. The edges and layout of Image 2 suggest it is a street scene with a clear view. This image does not have any empty space that matches the missing part in Image 1.

Step 3: Image 3 appears to show a person walking in front of a building. The scene aligns with the same architectural style as Image 1 (buildings with green walls), but the missing portion in Image 1 corresponds to the empty space in Image 1. The placement of Image 3 in this space would align with the architectural continuity.

Step 4: Considering the alignment of architectural elements and the absence of empty spaces in Image 2, Image 3 appears to be the correct choice to fill the missing portion in Image 1. Image 3 matches the scene and architectural style of Image 1.

Step 5: Since Image 3 fits the missing portion in Image 1, I conclude that Image 3 is the appropriate choice to place in the missing spot.

</think> <answer>the third image</answer>



### gpt-4o+CoT

"To determine which image fits the missing part of the first image, I need to observe the edges and the continuation of structures or colors. The first image has a missing lower right corner where part of a house is visible. Image 2 shows a continuation of the building in the correct position and color scheme, which seems to match the missing part of the first image. Image 3 does not align as well with the missing portion in terms of positioning and visual continuity. Therefore, the second image should fit the missing area, creating a seamless continuation from the first image to the second image."



Figure 6. Qualitative example 2. (1 of 2).



Image 1

Image 2

Image 3

**Question:** Given the first image with the lower right corner missing, can you tell which one of the second image or the third image is the missing part? Imagine which image would be more appropriate to place in the missing spot. You can also carefully observe and compare the edges of the images.  
**Answer Choices:** A. the second image B. the third image  
 Answer with the text of the option.

**gpt-5-thinking**

**Analyzing candidate images**  
 I need to find which of the two candidate images fits into the missing lower right corner of Image 1. This first image displays a scene of damaged buildings and rubble, likely from an earthquake, and I suspect the missing section shows a street with people.  
 Image 2 seems to show a similar scene, with debris and people present, possibly viewed from the adjacent area. Image 3 might showcase the right-hand segment of that same scene. I'll carefully analyze these options to determine the best fit.  
**Matching image edges**  
 I think the third image shows people running on the far right, possibly by a green shuttered door, which could match the edge continuity. I need to match these edges carefully along the border of the existing image and the missing section.  
 In Image 1, just to the right of the black box, I see rubble and a part of the green wall, along with a vertical pole that might be a lamp post. The missing part would likely display base-level rubble and the street. Image 2 has a vertical pole near center-right, and its left edge might continue the shape of the green wall and rubble.  
**Selecting image fit**  
 I notice that the left border of Image 2 shows a green wall, and the top of that image continues the roofline, which is a good sign. The horizon cloud might be similar too! There's also a vertical pole near the right side of Image 2 with a truncated horizontal bar.  
 In Image 3, the vertical pole is placed further right. I need to choose which image fits best in the lower right corner of Image 1. In Image 1, the black area covers the soldier's legs, while Image 2 has more people and debris in the center, versus Image 3, which has people running and possibly a door on the right edge.  
**Edge color matching analysis**  
 I'm focusing on the edge color matching for the missing section. The top edge of this area aligns with a horizontal line of rubble in the image, and I need the candidate piece to continue the same roofline and sky along its top edge. The black rectangle covers the lower right corner, meaning that top edge sits at mid-height within the image, overlapping with something above.  
 I'll need to find distinct objects at the vertical edge where the current image meets the missing area. I think there might be a person in a red shirt near that midline inside the black area.  
**Comparing images for alignment**  
 I realize that Images 2 and 3 show the same scene from different angles, so the correct missing piece should align with the upper part of the building in Image 1.  
 Looking at Image 2, the left side features green walls and a collapsed roof similar to Image 1, with a pile of rubble in the middle and people walking away. In Image 3, the scene shifts further right, and the missing lower right corner might include a vertical pole that's just visible above the black area in Image 1.  
 Above the black box in Image 1, there seems to be a tall cylindrical post next to the building, likely a utility pole with wires.  
**Evaluating vertical pole alignment**  
 In Image 2, I spotted a vertical pole on the right side, while Image 3 may have a pole closer to the left of center.  
 Hmm, another approach is comparing the skies: Image 1 shows a consistent sky with clouds, Image 2 has part cloud coverage and more buildings, while Image 3 shows less sky, possibly due to cropping.  
 When looking at the ground levels, Image 2 features a rubble cluster in the center, whereas Image 3 has a walkway on the right side with a man stepping. The boundary between the black area and visible portion reveals vertical parts of a green wall and collapsed roof on the left side, matching its appearance in Image 2.  
**Considering image placement**  
 If I place Image 2 in the missing lower right corner, its left border would need to align with what's immediately to the left of that missing area. This alignment is crucial for creating a seamless transition between the images. I want to ensure that any elements from both images, like walls, streets, or debris, match up nicely. This careful consideration of placement could really help in deciding which candidate fits best for that corner!

No answer provided due to thinking overlength.



Figure 7. Qualitative example 2. (2 of 3).



Image 1

Image 2

Image 3

**Question:** Given the first image with the lower right corner missing, can you tell which one of the second image or the third image is the missing part? Imagine which image would be more appropriate to place in the missing spot. You can also carefully observe and compare the edges of the images. **Answer Choices:** A. the second image B. the third image **Answer with the text of the option.**

### VigoRL

<think>The edge of Image 1 shows a building that has been partially destroyed, with debris scattered around. This matches the scene in Image 2 where there is a similar environment of destruction and debris. Thus, Image 2 seems like a plausible candidate for the missing part. (41, 97) Comparing the edges, the lower right corner of Image 1 has a dark area that aligns with the black square in Image 2, suggesting that Image 2 could be a continuation of the scene in Image 1. (150, 109) The lower right corner of Image 3 shows a person walking away from the camera, which is different from the scene of destruction in Image 1 and Image 2. Therefore, Image 3 does not seem appropriate to place in the missing spot. (200, 115) Considering the alignment and continuity of elements such as the road and buildings, Image 2 appears to be a continuation of the scene in Image 1, whereas Image 3 does not fit the context. (175, 100) Given the analysis, Image 2 matches the context and continuity better with Image 1, indicating that it would be more appropriate to place Image 2 in the missing spot. (180, 105)</think> <answer> the second image </answer>



### ViLASR

<think> To determine which image is most appropriate for the missing part, I need to compare the edges of the first image with the second and third images. The goal is to see which image, when placed in the missing part, aligns best with the rest of the image. Let's start by using the Object Mapper tool to map the edges of the images and see how they align. Action: Object Mapper [{"index": 1, "bbox\_2d": [210, 106, 379, 210], "label": "missing part"}] The visualization shows the missing part in Image 1. Now, I should use the Path Tracer tool to see how the edges of Image 2 and Image 3 align with the missing part. \n\nLet's start by tracing the path between the edges of Image 2 and the missing part. Action: Path Tracer [{"index": 2, "start\_point\_2d": [380, 106], "end\_point\_2d": [210, 106], "label": "from Image 2 to the missing part"}] The visualization shows the path from Image 2 to the missing part. Now, let's trace the path between the edges of Image 3 and the missing part to see which one fits better. Using the Path Tracer tool again to see how the edges of Image 3 align with the missing part. Action: Path Tracer [{"index": 3, "start\_point\_2d": [380, 106], "end\_point\_2d": [210, 106], "label": "from Image 3 to the missing part"}] The visualization shows the path from Image 3 to the missing part. Now, I should compare the visual paths and see which image, when placed in the missing spot, aligns best with the rest of the image. The visualization shows that Image 2 aligns better with the missing part compared to Image 3. The path tracer shows a clearer and more continuous flow between the images. Therefore, the second image is the most appropriate to place in the missing spot. Action: Answer <answer>A</answer>



Image 1

Image 2

Image 3

Image 1

Image 2

Image 3

Image 1

Image 2

Image 3

Figure 8. Qualitative example 2. (3 of 3).



Image 1



Image 2

**Question:** Question: Humans can find corresponding points for different objects in the same category. For instance, if there are images of two different cats, then the left ear tip of one cat corresponds to the left ear tip of the other cat, and the right front paw of one cat corresponds to the right front paw of the other cat. Given the following two images, a reference point is annotated on the first image, labeled with REF. You are given multiple red-circled points on the second image, choices of "A, B, C, D" are drawn beside each circle. Select between the choices on the second image and find the corresponding point for the reference point. Which point is corresponding to the reference point? Answer Choices: A. Point A B. Point B C. Point C D. Point D

### Thinking with Blueprint

"<think>I will first observe the reference point labeled "REF" in Image 1 because it serves as a starting point for identifying correspondences in Image 2. The bounding box coordinates for this object are [[550, 10], [580, 40]], and I will observe the object name as "REF" and its description as "A red-circled label labeled 'REF' positioned in the top-right corner of Image 1, near a monitor."

```
[{"object": "REF","bbox": [[550, 10], [580, 40]],"description": "A red-circled label labeled 'REF' positioned in the top-right corner of Image 1, near a monitor."}]
```

I have found all the objects necessary for reasoning. Now I need to do the reasoning based on the information I have obtained. Step 1: I observe Image 1 and identify the reference point labeled "REF" in the top-right corner. Its bounding box coordinates are [[550, 10], [580, 40]], and it is positioned near a monitor. Step 2: I move to Image 2 and examine the red-circled points labeled A, B, C, and D. I need to find a point that corresponds to the position of "REF" in Image 1. Step 3: I analyze Point A: It is located on the top-right corner of Image 2, above the television screen. This position closely aligns with the location of "REF" in Image 1. Step 4: I analyze Point B: It is positioned near the bottom of the television screen but slightly to the left of Point C. This placement does not align with the position of "REF." Step 5: I analyze Point C: It is positioned near the top-right corner of the television screen. This placement also aligns visually with the position of "REF." Step 6: I analyze Point D: It is positioned near the bottom of the television screen, below Point C. This placement aligns more with the bottom part of Image 1, where "REF" is not located. Step 7: Comparing all points, Point C in Image 2 aligns most closely with the position of "REF" in Image 1. Therefore, Point C is the corresponding point for "REF."

<answer>C</answer>



Figure 9. Qualitative example 3. (1 of 3).



Image 1



Image 2

**Question:** Question: Humans can find corresponding points for different objects in the same category. For instance, if there are images of two different cats, then the left ear tip of one cat corresponds to the left ear tip of the other cat, and the right front paw of one cat corresponds to the right front paw of the other cat. Given the following two images, a reference point is annotated on the first image, labeled with REF. You are given multiple red-circled points on the second image, choices of "A, B, C, D" are drawn beside each circle. Select between the choices on the second image and find the corresponding point for the reference point. Which point is corresponding to the reference point? Answer Choices: A. Point A B. Point B C. Point C D. Point D

gpt-4o+CoL

"First, identify the reference point labeled "REF" on Image 1, which is located at the top right corner of a monitor screen. The task is to find the corresponding point on Image 2. Examine Image 2, where multiple red-circled points labeled "A, B, C, D" are displayed. Look for a similar location on the screen as the reference point in Image 1. Point A, located on the top right corner of the screen in Image 2, corresponds to the reference point. The other points, B, C, and D, are positioned differently on the screen and do not match the location of the reference point. Thus, the corresponding point to "REF" in Image 1 is Point A on Image 2."

gpt-5-thinking

"Identifying reference points" I'm analyzing the reference point for the display. The REF tag is likely located at the top-right corner of the monitor, but is it on the outer edge or inside the display area? I need to be precise. In Image2, I want to see which option among A, B, C, and D matches that corner. It seems that if C is near the top-right inner corner of the display area, that would be the right one! Let's verify that visual alignment. Analyzing reference points" I'm looking at the reference points and trying to map them correctly. Point A is at the top-right corner of the entire device, which includes an outer white border. Meanwhile, C is in the top-right area of the black frame. Point D is in the bottom-left and B is at the bottom center. The REF point might correspond to the top-right corner of the glass. In Image1, the REF label is next to the dot at the display area's upper right. Let's determine which aligns with these! Confirming the correct answer" I remember that the REF dot is located at the top-right corner of the Mac's display, inside the screen, not on the device's frame. This aligns with my choice of point C."

No answer provided due to thinking overlength.

VigoRL

<think>\n\nThe reference point 'REF' in Image 1 seems to be related to the central object, which is a computer monitor. In Image 2, we need to find a similar central object. The central object in Image 2 appears to be the television screen. Point D is located near the center of the television screen, making it a candidate for the corresponding point. Let's consider Point B. Point B is located at the bottom-center of the television screen. This position is much closer to the central position of the reference point 'REF' in Image 1 than the previous points considered. Therefore, Point B seems to be the most promising candidate for the corresponding point.\n\n</think>\n\n<answer>Point B</answer>

Figure 10. Qualitative example 3. (2 of 3).



Image 1



Image 2

**Question:** Question: Humans can find corresponding points for different objects in the same category. For instance, if there are images of two different cats, then the left ear tip of one cat corresponds to the left ear tip of the other cat, and the right front paw of one cat corresponds to the right front paw of the other cat. In Given the following two images, a reference point is annotated on the first image, labeled with REF. You are given multiple red-circled points on the second image, choices of "A, B, C, D" are drawn beside each circle. Select between the choices on the second image and find the corresponding point for the reference point. Which point is corresponding to the reference point? Answer Choices: A. Point A B. Point B C. Point C D. Point D

### ViLASR

<think> To determine the corresponding point, I should first identify the reference point on the first image and then find a similar point on the second image. The reference point is labeled as REF on the first image. Now, I need to use the Path Tracer tool to see if any of the points A, B, C, or D on the second image correspond to the reference point. Let's start by using the Path Tracer tool to trace the reference point on the first image to the points on the second image.

Action: Path Tracer [{"index": 1, "start\_point\_2d": [370, 20], "end\_point\_2d": [498, 230], "label": "trace from reference point to point D"}]

The visualization shows the path tracer connecting the reference point (REF) on the first image to the point D on the second image. The path tracer indicates that point D on the second image is the corresponding point for the reference point. Now, let's use the Object Mapper tool to confirm the corresponding objects and see if the path tracer aligns with our reasoning.

Action: Object Mapper [{"index": 2, "bbox\_2d": [498, 230, 514, 246], "label": "reference point"}]

The visualization with the Object Mapper tool confirms the reference point on the second image. The reference point is correctly mapped to the point D on the second image. Given the visual confirmation, the corresponding point for the reference point is indeed point D. The other points A, B, and C do not align with the reference point. Thus, the final answer is D. </think>

Action: Answer <answer>D</answer>



Image 1

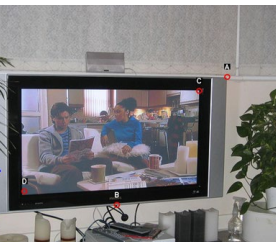


Image 2



Image 1



Image 2



Figure 11. Qualitative example 3. (3 of 3).



**Question:** A point is circled on the first image, labeled with REF. We change the camera position or lighting and shoot the second image. You are given multiple red-circled points on the second image, choices of "A, B, C, D" are drawn beside each circle. Which point on the second image corresponds to the point in the first image? Answer Choices: A. Point A B. Point B C. Point C D. Point D Answer with the text of the option.

### Thinking with Blueprint

"<think> I will first observe the red-circled point labeled "REF" in Image 1 because it corresponds to the question's requirement and is explicitly mentioned as a reference point. I will focus on this object in Image 1. The bounding box coordinates for this object are [[300, 100], [350, 150]], and I will update the JSON accordingly.

```
{ "object": "REF marker", "bbox": [[300, 100], [350, 150]], "description": "A red-circled marker labeled 'REF' near a construction area in Image 1." }
```

I have found all the objects necessary for reasoning. Now I need to do the reasoning based on the information I have obtained. Step 1: I observe Image 1, where a red-circled marker labeled "REF" is visible near the construction area on the right-hand side of the image. The bounding box coordinates for this marker are [[300, 100], [350, 150]], indicating its position relative to the statue and the surrounding area.

Step 2: I compare Image 2 to Image 1. I observe that Image 2 contains multiple red-circled points labeled "A," "B," "C," and "D." I need to identify which of these corresponds to the "REF" marker in Image 1.

Step 3: I analyze the spatial arrangement of the red-circled markers in Image 2:

- Point A is positioned near the head of the statue, slightly towards the top region.
- Point B is positioned near the base of the statue, close to the ground and slightly towards the bottom-right corner.
- Point C is positioned further down and to the right of the statue's base.
- Point D is positioned near the construction area beside the statue.

Step 4: Since "REF" in Image 1 is located near the construction area and slightly below the statue's head, I compare this location to Image 2. Point D aligns most closely with the position of "REF," as it is near the construction area and aligns spatially with Image 1.

Step 5: I confirm that Point D in Image 2 corresponds to the "REF" marker in Image 1. The bounding box [[300, 100], [350, 150]] in Image 1 matches Point D's position in Image 2.

Step 6: I finalize the selection of the correct answer choice based on the observations and bounding box alignment. The answer is Point D.

<answer>Point D</answer>



Figure 12. Qualitative example 4. (1 of 3).



**Question:** A point is circled on the first image, labeled with REF. We change the camera position or lighting and shoot the second image. You are given multiple red-circled points on the second image, choices of "A, B, C, D" are drawn beside each circle. Which point on the second image corresponds to the point in the first image? Answer Choices: A. Point A B. Point B C. Point C D. Point D Answer with the text of the option.

### gpt-4o+CoT

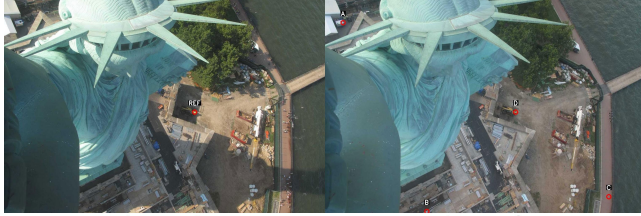
"First, identify the position of the red-circled point labeled 'REF' in Image 1. The point is located among an area of construction and pathways surrounding the Statue of Liberty. Now, compare the positions of the red-circled points labeled 'A, B, C, D' in Image 2. Look for distinctive landmarks or features around the REF point in Image 1 and find a matching setup in Image 2. It appears that the arrangement around Point A in Image 2 is most similar to the surroundings of the REF point in Image 1, including the pathways and construction layout. Thus, Point A corresponds to REF." ❌

### gpt-5-thinking

\*\*Analyzing image orientation\*\*\n\nI'm trying to figure out the placement of the red dot above the walkway boundary. I want to mentally zoom in and compare the camera's orientation, even though the region remains the same. In Image 2, point A is near the statue's head at the top left, but options A, B, C, and D are lower down. I need to carefully observe these images again. The first half of Image 1 has the green statue's head, and the REF is near a black elongated shape. \*\*Identifying features and points\*\*\n\nI'm examining the features around the statue, including a triangular walkway and a black line near its right side that leads to an area with pipes, white materials, and containers. There's also a rectangular patch of ground near the water wall. I'm trying to determine where REF is, possibly near a cluster of small white cylinders. In the second image, point A is near the statue's crown but seems too far from REF. Points B and C are closer to the statue and walkway. I need to find proximity to the tree area. \*\*Analyzing image features\*\*\n\nIn image 2, I see a tree cluster at the top right, adjacent to the walkway. REF in image 1 is near that tree cluster, so I'm thinking the corresponding point in image 2 would be nearby. Of the points A, B, C, D, only D might be near the trees; it's labeled near the middle right close to the walkway. Point A is at the top left near the crown, likely not near the trees. I need to analyze the second image more closely to confirm. \*\*Determining point locations\*\*\n\nI'm looking at the circles in the second image, where 'A' is near the top left, 'B' is at the bottom left, 'C' is in the bottom right, and 'D' is near the mid-right by the trees. I think point D seems plausible because it corresponds with the placement of REF in image 1, which is close to a small orange object near the walkway and tree cluster. Given the positions, I'm leaning towards point D as the answer. \*\*Verifying point selection\*\*\n\nI'm double-checking the situation with the black rectangular trench or walkway that goes diagonally from the statue to the bottom right. In image 1, this walkway is present, and next to it is a rectangular excavation area where the red dot might be. In image 2, it seems point D aligns with that area. I need to return the exact text of the answer choice, which is 'Point D.' There's no need to check for image safety since it's a public structure without humans. Let's finalize with 'Point D.' ❌

No answer provided due to thinking overlength.

Figure 13. Qualitative example 4. (2 of 3).



**Question:** A point is circled on the first image, labeled with REF. We change the camera position or lighting and shoot the second image. You are given multiple red-circled points on the second image, choices of "A, B, C, D" are drawn beside each circle. Which point on the second image corresponds to the point in the first image? Answer Choices: A. Point A B. Point B C. Point C D. Point D Answer with the text of the option.

### VigoRL

<think>\nTo determine which point corresponds to \"REF\" in the first image, I'll compare the position of \"REF\" with the surrounding elements like the statue's head and the water. In the second image, point A is located near the top of the statue's head, which aligns with the position of \"REF\" in the first image. Let's examine point A further. It is situated close to the statue's head, just like \"REF\" is, suggesting that point A might be the correct match. Next, I will analyze point B. Point B is positioned lower down on the statue, closer to the pedestal. This placement is quite different from the position of \"REF,\" which is near the top of the statue's head. Therefore, point B is unlikely to be the correct match. Now let's consider point C. Point C is located on the ground, far from the statue and water. Its position does not match the location of \"REF,\" which is near the top of the statue's head. Thus, point C seems unrelated to \"REF.\" The previous analysis indicates that point A is the only candidate matching the position of \"REF.\" Let's also check point D. Point D is near the statue's face, but it is not directly above the statue's head like \"REF.\" The placement of point D suggests it is not the correct match for \"REF.\" Having verified all the points, it is clear that point A is the only one that aligns with the position of \"REF\" in the first image.\n</think>



### ViLASR

<think> To determine which point on the second image corresponds to the point in the first image, I need to establish a connection between the two images. The point labeled \"REF\" in the first image should be compared with the red-circled points in the second image. I can use the Path Tracer tool to draw a connection between the \"REF\" point and the corresponding point in the second image. Let's start by using the Path Tracer tool to see which point in the second image aligns with the \"REF\" point in the first image.

Action: Path Tracer [{"index": 1, "start\_point\_2d": [260, 124], "end\_point\_2d": [260, 124], "label": "Connection between images"}]

The path tracer tool indicates that the point labeled \"REF\" in the first image corresponds to the point labeled \"A\" in the second image. The visualization confirms that the point \"REF\" aligns with point A, and no other points in the second image match the position of \"REF\". Let's confirm the final answer using the information from the visualization.</think>

Action: Answer\n<answer>A</answer>



Figure 14. Qualitative example 4. (3 of 3).



Image 1

Image 2

**Question:** How did the camera likely rotate when shooting the video?  
**Answer Choices:**  
 A. rotated right  
 B. rotated left  
**Answer with the text of the option.**

**Thinking with Blueprint**

"<think>I will focus on the position of the computer monitor, glass cabinets, and the overall spatial orientation of the scene in both images. The computer monitor is clearly visible on the right side of Image 1 and appears more prominently on the left side of Image 2. The glass cabinets and boxes are positioned to the left of the monitor in Image 1 but shift towards the center in Image 2.

I will focus initially on the computer monitor in Image 1 as it provides a clear reference point for spatial orientation and its shifting position will help infer camera rotation. The monitor is positioned on the right side of the scene in Image 1 and is more clearly visible in Image 2 on the left side, indicating a rotational motion. The bounding box coordinates for the computer monitor in Image 1 will be positioned in the right foreground of the scene.

```
["object": "computer monitor",\n  "bbox": [[700, 300], [900, 500]],\n  "description": "A rectangular black computer monitor positioned on the right foreground of Image 1."]
```

I have found all the objects necessary for reasoning. Now I need to do the reasoning based on the JSON and the image. I will now reason step by step. The JSON indicates the computer monitor's position shifts from the right foreground in Image 1 to the left foreground in Image 2, suggesting a rotational movement of the camera. The scene's perspective also shifts such that the glass cabinets, previously positioned on the left, now appear more centered.\n\nThe camera's perspective in Image 1 is oriented such that the computer monitor is clearly on the right. However, in Image 2, the computer monitor has shifted to the left of the frame, and the glass cabinets and boxes on the left become more prominent. This indicates a leftward rotation of the camera.

Hence, the camera rotated to the left when transitioning between the two images.\n</think>“  
 <answer>rotated left</answer>



**gpt-4o+CoT**

To determine the direction of camera rotation between Image 1 and Image 2, it's essential to analyze how the scene shifts from one image to another. Image 1 features more space to the left (a view of shelving and a workstation), while Image 2 shows the space that was to the right (more direct view of the boxes and glass door). This indicates the camera rotated towards the right from Image 1 to Image 2.



Figure 15. Qualitative example 5. (1 of 3).



Image 1

Image 2

**Question:** How did the camera likely rotate when shooting the video?  
**Answer Choices:**  
 A. rotated right  
 B. rotated left  
**Answer with the text of the option.**

**gpt-5-thinking**

**Analyzing camera rotation** need to answer the question about how the camera likely rotated when shooting the video by looking at the two images labeled Image 1 and Image 2. I should analyze the positions of objects in each frame to determine the rotation direction. Image 1 appears to be in a portrait orientation, while it shows the left side of a lab with monitors and cabinets. Comparing it to Image 2, which shows the center area, suggests that the camera rotated either to the right or the left.  
**Determining camera rotation direction** need to identify if the second view is to the right relative to the first. In Image 1, there's a tall column near the center-right, with cabinets behind it and possibly a box stack to its right. In Image 2, the same column is near the right edge, and the scene seems shifted, revealing more of a glass door in the center. This shift suggests the camera may have rotated left. I should compare references like the column and overhead beams to clarify the motion more clearly.  
**Analyzing camera pan direction** I'm considering whether the camera panned to the right, moving from the monitors towards the glass door, which is positioned to the right. To visualize this, I imagine standing at a desk facing the room. If I rotate the camera right, objects on the right would move into the center view. In Image 1, the glass wall appears in the background near the column, while in Image 2, it's central. The monitors, previously visible to the left, are now out of view, suggesting that the camera rotation was indeed to the right. Still, I want to be completely certain.  
**Clarifying camera rotation** I'm looking closely at the shift of boxes near the column. In Image 1, the boxes are on the far right, while in Image 2, they are more centered. This suggests the camera likely rotated to the right. I need to answer with the exact text option, which is "rotated right." But I should consider—could it actually be left? If the monitors are on the left and the second image shows more to the right, wouldn't that mean turning left? I realize that if the objects move left in the frame, that corresponds with a right rotation. So by analyzing the optical flow, if the scene changes from left to right, the camera likely rotated left. It's a bit tricky, but I'm leaning towards "rotated right" based on the boxes. So, the answer is "rotated right." ❌

**VigoRL**

**To determine how the camera rotated,** I'll compare the position of objects in both images. In Image 1, the desk with monitors is more towards the left side, while in Image 2, it has moved closer to the center. This suggests a rotation towards the right. (350, 300) Additionally, the shelving units appear more aligned with the edge of the frame in Image 1 compared to their placement in Image 2, indicating that the camera has indeed rotated to the right. (150, 200) The glass partition that runs vertically in the middle of Image 1 is no longer present in Image 2; instead, we see more of the wall space. This further supports the idea that the camera has rotated to the right. (600, 250) In Image 1, the window is visible on the far right, whereas in Image 2, it is more central. This shift in perspective aligns with the camera rotating to the right. (850, 300) The door in Image 1 is located towards the bottom right corner, but in Image 2, it is more centrally positioned. This indicates the camera has rotated to the right, bringing the door into the center. (900, 400) The large cardboard boxes in Image 1 are on the left side, but in Image 2, they are more towards the center-right. This rotation effect continues to support the conclusion that the camera has rotated right. (120, 350) The overall layout and positioning of all elements suggest a consistent rotation pattern across various objects. Considering the consistent changes in object positions, the camera has indeed rotated right. (400, 450)

**answer** rotated right  
**answer**

Figure 16. Qualitative example 5. (2 of 3).



Image 1

Image 2

**Question:** How did the camera likely rotate when shooting the video?  
**Answer Choices:**  
 A. rotated right  
 B. rotated left  
**Answer with the text of the option.**

**ViLASR**

<think> To determine the direction of the camera's rotation, I should first identify key objects and their relative positions in the images. The Object Mapper tool will help me pinpoint these objects, and their relative positions will give clues about the camera's movement. After that, I can use the Path Tracer tool to visualize the potential path of the camera's rotation.

Let's start by identifying the objects in the image using the Object Mapper tool.  
 Action: Object Mapper [{"index": 1, "bbox\_2d": [225, 240, 372, 316], "label": "desk in Image 1"}],

The visualization shows the desk in Image 1, and now I need to find the desk in Image 2 to understand the relative change. The Path Tracer tool will help me visualize the potential path of the camera's rotation.

Let's use the Object Mapper tool again to identify the desk in Image 2.  
 Action: Object Mapper [{"index": 2, "bbox\_2d": [812, 232, 899, 320], "label": "desk in Image 2"}],

Now that we have the relative positions of the desks in both images, we can use the Path Tracer tool to visualize the potential path of the camera's rotation. This will help us understand the direction of the camera's movement.

Let's use the Path Tracer tool to visualize the potential path of the camera's rotation.  
 Action: Path Tracer [{"index": 1, "start\_point\_2d": [294, 274], "end\_point\_2d": [854, 274], "label": "potential camera rotation path"}]

The visualization shows the potential path of the camera's rotation, but the path tracer shows a straight line, which may not accurately represent the camera's movement. The camera likely rotated around a central point, not following a straight path. Let's re-evaluate the relative positions and consider the likely central point of rotation. The camera probably rotated around the central area of the two images, which would suggest a leftward rotation. The camera likely rotated left.

Answer  
 <answer>B</answer>



Figure 17. Qualitative example 5. (3 of 3).



**Question:** *Is the tv behind or in front of the red stool?*

**Answer Choices:**

A. *Behind*

B. *In front*

**Answer with the text of the option.**

### Thinking with Blueprint

<think>

I will first focus on the red stool because it is a prominent foreground object in the image. The red stool is clearly distinguishable due to its bright red color and its placement closer to the bottom-right foreground of the image. The bounding box coordinates for the red stool will be located prominently in the foreground.

```
[{"object": "red stool", "bbox": [[400, 400], [500, 500]], "description": "A bright red-colored stool situated in the bottom-right foreground of the image."}]
```

I will then focus on the tv, situated clearly in the top right background of the image. The bounding box coordinates for the TV will be located top right in the background.

```
[{"object": "red stool", "bbox": [[400, 400], [500, 500]], "description": "A bright red-colored stool situated in the bottom-right foreground of the image."}, {"object": "TV", "bbox": [[410, 0], [500, 100]], "description": "A bright red-colored stool situated in the bottom-right foreground of the image."}]
```

I have found all the objects necessary for reasoning. Now I need to do the reasoning based on the JSON and the image. I will determine the spatial relationship between the red stool and the TV by analyzing their relative positions in the image. The red stool is clearly positioned closer to the foreground of the image, while the TV is located further away, in the midground behind a window. The red stool's location is clearly to the foreground, and the TV is positioned behind the red stool, based on the spatial arrangement.

Step-by-step reasoning:

1. I identified the red stool in the foreground, located in the bottom-right foreground of the image.
2. I identified the TV positioned further back, behind a window, and higher in the image.
3. Comparing the spatial arrangement, the TV is clearly further away from the viewer and behind the red stool.

Hence, the TV is behind the red stool.

</think>

<answer>Behind</answer>



Figure 18. Qualitative example 6. (1 of 2).



**Question:** Is the tv behind or in front of the red stool?

**Answer Choices:**

A. Behind

B. In front

**Answer with the text of the option.**

### gpt-4o+CoT

"The image shows a dental office setting. There is a red stool towards the right, and the television is mounted near the window, behind the stool." Answer: In front



### gpt-5-thinking

**\*\*Determining spatial relation\*\*** I need to answer whether the TV is behind or in front of the red stool in the image of the dental office. The red stool is located in the foreground while the TV is mounted on the wall in the background. Since the stool is closer to the camera and the TV is further back, it's clear that the TV is behind the stool. So, I'll choose the option "behind" and confirm that answer. (The final answer reported is in front.)



### VigoRL

<think>\n\nTo determine if the TV is behind or in front of the red stool, I need to compare their positions relative to each other. The TV is located at (230, 20) and the red stool is at (180, 230). The TV is positioned higher up in the image compared to the red stool, indicating that it is behind the stool. Let's verify this by considering another point. The TV is also connected to wires which run down towards the stool, suggesting that the TV is behind the stool. Another verification point is the position of the monitor at (240, 45) which is clearly behind the stool. This further supports that the TV is behind the stool. Finally, let's consider the position of the chair behind the stool. The chair back is located at (150, 150) and is also behind the stool, which confirms that the TV is indeed behind the stool. All these points consistently show that the TV is behind the stool.\n</think>\n<answer>behind</answer>



### ViLASR

<think> To determine the relative position of the TV, I should first identify the red stool and the TV in the image. The TV is likely located near the upper right of the image, and the red stool can be seen in the lower right. The relative positioning can be better understood by using the Path Tracer tool to visualize the relative position of the TV concerning the red stool. Action: Object Mapper [{"index": 1, "bbox\_2d": [194, 214, 252, 248], "label": "red stool"}]\n\nThe red stool's position is correctly identified. Now, let's locate the TV in the image, which is likely near the upper right corner. After identifying the TV, we can use the Path Tracer tool to see its relative position concerning the red stool. Action: Object Mapper [{"index": 2, "bbox\_2d": [208, 26, 262, 52], "label": "TV"}]\n\nThe TV and the red stool are correctly identified. Now, I will use the Path Tracer tool to see the relative position of the TV concerning the red stool. The path tracer will help visualize the relative positioning clearly. Action: Path Tracer [{"index": 3, "start\_point\_2d": [227, 229], "end\_point\_2d": [230, 32], "label": "from red stool to TV"}]\n\nThe visualization with the path tracer shows the relative position of the TV concerning the red stool. The path tracer indicates that the TV is behind the red stool. The relative positioning is clear now.\n</think>\nAction: Answer\n<answer>A</answer>

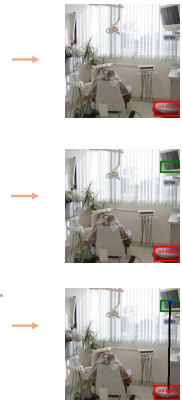


Figure 19. Qualitative example 6. (2 of 2). We can observe inconsistency between reasoning and answer in both gpt-4o + CoT and gpt-5-thinking.



**Question:** Which way is the person turning from his perspective?

Answer Choices:

A. Left

B. Right

Answer with the text of the option.

### Thinking with Blueprint

<think>

I will focus on the person skiing in the foreground of the image to understand their orientation and the direction they appear to be turning. The person is clearly visible on the left side of the image, positioned on a snowy trail. The bounding box coordinates for the person will be located in the left-center foreground of the image.

The first object to observe will be the skier in the foreground of the image, positioned on the left side of the scene, wearing dark clothing and standing on a snowy trail. This person is clearly distinguishable due to their position and the context of the image. The bounding coordinates for the skier will be in the left-center foreground of the image.

[{"object": "person", "bbox": [[200, 200], [300, 400]], "description": "A skier positioned on the left side of the image, standing on a snowy trail."}]

I have found all the objects necessary for reasoning. Now I need to do the reasoning based on the JSON and the image. From the JSON, the skier is positioned on the left side of the image, standing on a snowy trail. Observing the direction of the skier's turn, the person is facing towards the camera and appears to be turning towards the right side of the scene in the camera's perspective. The skier's orientation aligns with the direction of the trail's curves in the image. As the skier's left and right is the opposite of the camera's, the skier is turning left.

I will now confirm the direction of the person's turn based on their position and the context of the scene.

</think>

<answer>Left</answer>



### VigoRL

<think>\n\nTo determine which way the person is turning, I need to observe the direction of their body and skis relative to the path they've made. The skis are angled towards the right side of the image (479, 308), suggesting they are turning in that direction. Next, I'll consider the person's body orientation. Their body is also angled slightly to the right, indicating they are leaning into the turn (375, 226). To confirm this observation, I'll check if there are any objects or features in the image that suggest an obstacle or a change in terrain that would necessitate a left turn. There are no such features visible on the left side of the image (150, 250). This further supports the conclusion that the person is turning right. Finally, considering the overall context and the absence of any indication of a left turn, such as a change in the direction of the tracks or any visual cues suggesting a left turn (250, 400), the correct answer is that the person is turning right.\n\n</think><answer>Right</answer>



Figure 20. Qualitative example 7. (1 of 2).



**Question:** Which way is the person turning from his perspective?

**Answer Choices:**

A. Left

B. Right

**Answer with the text of the option.**

gpt-4o+CoT

"The skier is shown moving on a snowy path. The tracks on the snow indicate a curve. From the skier's perspective, he appears to be turning slightly to his left. So the correct answer is based on identifying the direction of the curve from the skier's perspective." Answer: Left



gpt-5-thinking

**Analyzing ski tracks**  
 Let's focus on the ski tracks near the person. There are two tracks left of the main trail, which curve to the left. It seems the person could be turning left since his skis angle that way, but I have to be careful with perspective. If he's facing the camera, then his left is on our right. So, if he's leaning toward the viewer's right, that seems to match his torso angle. His legs may curve differently.  
**Evaluating ski track direction**  
 The ski tracks curve to the left in the image, which translates to the person's right if he's facing us. This suggests he's turning right. The main path slopes from the bottom right to the upper center, with the person likely on the left track. I notice his skis point toward the lower left, and the tracks behind him curve. There's confusion about his body lean; it seems his upper body is leaning to his right to maintain his direction. I'm puzzling it out!  
**Determining turning direction**  
 I'm trying to clarify which way the person is turning. If he's turning left from his perspective, his skis would lean that way, and the tracks behind indeed show a leftward curve. However, for us watching, it appears he's turning left, but from his perspective, that would actually be right. Many exam datasets have tricky questions like this. After reviewing, it looks like he's indeed turning to our left, meaning he's turning right from his perspective. So, I'll go with "right" as my answer choice.  
**Examining pole orientation**  
 It's tricky figuring out the pole orientation. If the person is facing away from us, then he would be turning to his left, which means we'd see it as a left turn. Now, to determine which way is more plausible, I can look at the details, like his elbows or maybe the front of his jacket or backpack. Since those aren't clear, I'll focus on the pole angles. If the poles are behind him relative to the viewer, that gives me some clues.

No answer provided due to thinking overlength.



VILASR

"<think> To determine the direction of the turn, I need to visualize the person's perspective and see which way they are turning relative to their path. I'll use the Path Tracer tool to see the direction of the turn more clearly. The person is turning to the right from their perspective.

Action: Path Tracer [{"index": 1, "start\_point\_2d": [147, 210], "end\_point\_2d": [160, 255], "label": "turn direction"}],

The path tracer visualization confirms the person is turning to the right. The path tracer shows the direction of the turn, and it aligns with the right side of the image. The person is turning right from their perspective.

</think>

Action: Answer\n<answer>B</answer>



Figure 21. Qualitative example 7. (2 of 2).

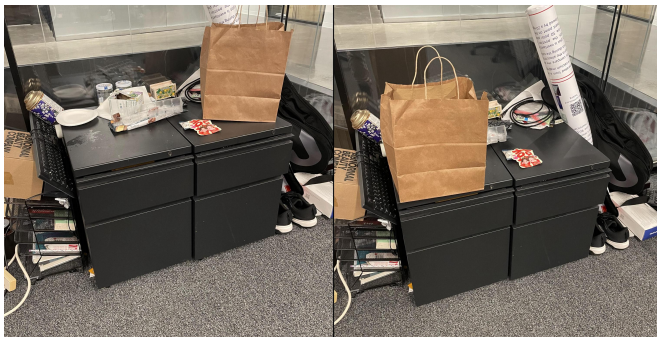


Image 1

Image 2

**Question:** Were any of the objects in the initial frame that you can still see in the second frame moved from their original positions?

**Answer Choices:**

A. paper bag was moved left and towards the camera  
 B. paper bag was moved right and towards the camera

Answer with the text of the option.

### Thinking with Blueprint

<think>I will first focus on the paper bag because it is clearly visible and distinct on the right-hand side of the images. The bag is positioned on top of the black furniture, making it highly noticeable. I will first observe the paper bag because it is clearly positioned on the right-hand side of the image atop the black furniture. It is easily distinguishable by its brown color and location. The paper bag is an essential object in this scene as it is clearly visible and its movement is a key point of interest in the question. The bounding box coordinates for the paper bag are located in the upper of both images. The first object I would focus on is the paper bag, as it is easily recognizable by its brown color and position atop the black furniture. It is clearly positioned in the upper central-right of the image, making it highly distinguishable. The coordinates for the paper bag will be roughly located in the upper right quadrant of the image. [{"object": "paper bag 1", "bbox": [[670, 110], [870, 300]], "description": "A brown paper bag positioned on top of the black furniture in the upper central-right of the image."}]

I will then focus on the "paper bag" because it is clearly positioned on the middle left of the image atop the black furniture. It is highly distinguishable by its brown color and location in the scene. The bounding box coordinates for the paper bag are likely situated in the middle left of the image.

[{"object": "paper bag 1", "bbox": [[670, 110], [870, 300]], "description": "A brown paper bag positioned on top of the black furniture in the upper central-right of the image."}]

{"object": "paper bag 2", "bbox": [[150, 210], [370, 520]], "description": "A brown paper bag positioned on top of the black furniture in the middle left of the image." } ` ` `

I have found all the objects necessary for reasoning. Now I need to do the reasoning based on the JSON and the image. I will analyze the movement of the "paper bag" between the two images. The bag's position is clearly on the upper central-right of the furniture in both images. In Image 1, the paper bag appears further to the left, while in Image 2, it seems closer to the camera and slightly shifted to the right. By comparing the bounding box coordinates and the relative positions of the furniture and the bag, the paper bag appears to have moved slightly left and closer to the camera in Image 2. I will now conclude based on the options provided.

</think>

<answer> paper bag was moved left and towards the camera</answer>



Figure 22. Qualitative example 8. (1 of 3).

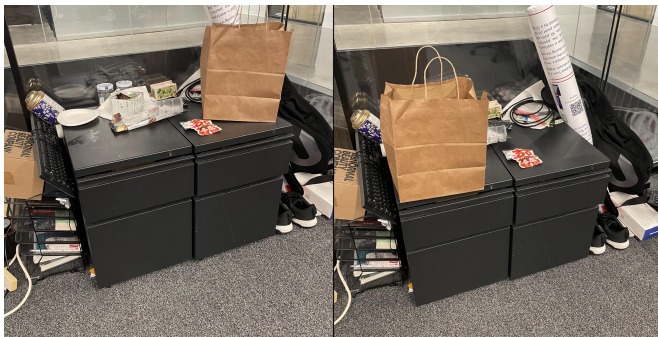


Image 1

Image 2

**Question:** *Were any of the objects in the initial frame that you can still see in the second frame moved from their original positions?*

**Answer Choices:**

*A. paper bag was moved left and towards the camera  
B. paper bag was moved right and towards the camera*

*Answer with the text of the option.*

### gpt-4o+CoT

"To determine whether any objects were moved between the two images, I need to compare the position of the paper bag in Image 1 and Image 2. In Image 1, the paper bag is placed more towards the left and further back on the drawer. In Image 2, the paper bag appears to have been moved to the right and closer to the front of the drawer. Based on the available options, this corresponds to the second choice.



### gpt-5-thinking

**\*\*Analyzing bag movement between images\*\***  
 need to decide the movement of the paper bag between two images. In the first image, the bag is sitting more to the right on the top of the left pedestal, whereas in the second image, it appears to have moved left. I notice that in image one, the bag is positioned on the right side of the left cabinet, while in the second, it seems to be closer to the camera and also more centered to the left. It's tricky to gauge the exact distances, though!  
**\*\*Confirming bag movement direction\*\***  
 In image 2, the bag is sitting closer to the front edge and slightly overlapping the seam, angled in such a way. I have only two options for movement: either "the paper bag was moved left and towards the camera," or "the paper bag was moved right and towards the camera." Based on my impression, the bag indeed moved leftward and forward. To confirm, the bag's position in image 1 is nearer to the right edge, while in image 2, it's clearly leftward. So, the answer is: "The paper bag was moved right and towards the camera."



### VigoRL

<think> To determine if the paper bag has been moved, I'll compare its position in both images. In Image 1, the paper bag is positioned closer to the center-left side of the image at (350, 100). In Image 2, the paper bag appears slightly further right and closer to the camera at (850, 150). This indicates that the paper bag was indeed moved. Now let's consider the direction of movement. Comparing the positions of the paper bag in both images, it has shifted to the right compared to its original position at (350, 100) in Image 1. It has also moved towards the camera, as it appears larger in Image 2 at (850, 150). This aligns with choice B. We need to verify if there's any other possibility. The paper bag's location has been confirmed to shift right and towards the camera, which makes option A unlikely. Let's check if the bag was moved left. However, the bag is clearly positioned more towards the right in Image 2, so this possibility can be ruled out. Lastly, let's consider if the bag was moved in another direction. Since the bag has been confirmed to move right and towards the camera, no other directions match the observed movement. Therefore, we can conclude that the paper bag has been moved right and towards the camera. </think> <answer>paper bag was moved right and towards the camera</answer>



Figure 23. Qualitative example 8. (2 of 3).



Image 1

Image 2

**Question:** Were any of the objects in the initial frame that you can still see in the second frame moved from their original positions?

**Answer Choices:**

A. paper bag was moved left and towards the camera  
 B. paper bag was moved right and towards the camera

Answer with the text of the option.

### VILASR

<think>"<think> To determine if any objects were moved, I should first identify the objects in the initial frame and see if they appear in the same position in the second frame. The paper bag is a key object to focus on, as its movement would be most noticeable. I will use the Object Mapper tool to map the paper bag in the initial frame and see where it appears in the second frame.

Action: Object Mapper [{"index": 1, "bbox\_2d": [250, 18, 340, 91], "label": "paper bag"}]

The visualization confirms the paper bag's position in the initial frame. Now, I should use the Path Tracer tool to see if there's any movement of the paper bag in the second frame. This will help determine if the paper bag was moved and in what direction.

Action: Path Tracer [{"index": 2, "start\_point\_2d": [290, 50], "end\_point\_2d": [560, 160], "label": "movement of the paper bag"}]

The visualization shows the path tracer indicating the movement of the paper bag from the left side of the image to the right side. The paper bag was indeed moved, and the path tracer shows the direction of movement. Now, I need to confirm the relative movement in terms of left or right and towards the camera or away from the camera.</think>

Action: Answer <answer>B</answer>"



Image 1

Image 2

Image 1

Image 2

Figure 24. Qualitative example 8. (3 of 3).



**Question:** *If I sit by the dog and face the same direction as the dog, will the lamp be to my left or right?*

**Answer Choices:**

A. *Left*

B. *Right*

*Answer with the text of the option.*

### Thinking with Blueprint

<think>I will focus on the lamp as the first object to identify its position relative to the person sitting next to the dog. The lamp is located on a white wall behind the couch and is positioned to the left side of the image.\n\nI will first focus on the lamp, which is positioned on the white wall near the top-left corner of the image, as it is the most distinctive object for spatial orientation. The lamp is a cylindrical object standing upright near the window. The bounding box coordinates for the lamp will be located in the upper-left region of the image.

```
[{"object": "lamp", "bbox": [[290, 20], [360, 300]], "description": "A cylindrical lamp standing upright near the window on the white wall."}]
```

I will then focus on the dog, as it is clearly positioned in the center of the image. The dog is lying on the sofa, facing left.

```
[{"object": "lamp", "bbox": [[290, 20], [360, 300]], "description": "A cylindrical lamp standing upright near the window on the white wall."}]
```

```
{"object": "dog", "bbox": [[400, 400], [560, 460]], "description": "A dog lying on the sofa."}]
```

I have found all the objects necessary for reasoning. Now I need to do the reasoning based on the JSON and the image. The lamp, identified in the JSON with bounding coordinates near the top-left corner of the image, is positioned behind the couch and to the left relative to the person sitting next to the dog.

If a person were to sit by the dog and face the same direction as the dog, their perspective would align with the dog's view, which is oriented towards the right side of the image (towards the couch). The lamp, therefore, would be positioned to the left of the person's perspective.

Hence, the lamp would be on the person's left if they were to sit by the dog and face the same direction.</think>

<answer>right</answer>



### gpt-4o+CoT

To determine the direction of the lamp from the perspective of someone sitting by the dog and facing the same direction, observe the image. The dog is facing towards the right side of the sofa. The lamp is positioned behind the dog on the left side of the picture. If you sit by the dog and face the same direction as the dog, the lamp will be to your left.



Figure 25. Qualitative example 9. (1 of 3).



**Question:** *If I sit by the dog and face the same direction as the dog, will the lamp be to my left or right?*

**Answer Choices:**

A. Left

B. Right

**Answer with the text of the option.**

gpt-5-thinking

**Determining lamp position**  
The question is about figuring out the lamp's position if I sit by the dog facing the same direction as it. First, I need to understand the dog's orientation. It looks like the dog's head is facing left, but it also seems to point towards the viewer. Then there's the lamp, which I see is positioned behind the couch on the left side of the image. So, if I'm sitting next to the dog, the lamp would be to my left.  
**Clarifying lamp orientation**  
I need to clarify the lamp's position relative to the dog's direction. If I'm facing the same way as the dog, which is towards the camera, my left would be on the viewer's right side. The lamp is positioned behind the couch, which is to the left side of the viewer. So, if I'm sitting next to the dog and facing forward like it, the lamp would be behind and to my left based on that orientation.  
**Assessing dog's orientation**  
I'm reconsidering the dog's orientation in the image, thinking it might actually be facing the right side of the picture, towards the sofa arm. If the dog's head is near the right edge with its nose pointing right, then it is indeed facing that way. So if I sit by the dog facing right, it would mean the lamp is behind to my left. I'm trying to analyze the dog's position based on its body and paw placement, which seems a little tricky!  
**Analyzing lamp's position**  
I'm examining the lamp's position relative to the dog. The lamp is positioned behind the couch, likely on the left side of the dog. From my perspective, if I sit next to the dog facing the same direction as it, the lamp would indeed be to my left. So, from the dog's viewpoint, if it's facing right, the lamp is behind to the left, which I need to clarify: it's definitely to the left of the dog.  
**Determining lamp relative to dog**  
I'm working on figuring out the lamp's position relative to the dog's head. The lamp is indeed to the left of the dog's head, which is on the right side of the image. The question asks if the lamp will be to my left or right if I sit by the dog facing the same direction. Generally, they assume the dog is facing forward; if that's the case, the lamp is behind and to the left of the dog. People often choose "left" based on this orientation.

✗

VigoRL

`<think>`  
To determine if the lamp would be to the left or right if sitting by the dog facing the same direction, I need to consider the dog's position relative to the lamp. The dog is lying on the couch facing towards the bottom right corner of the image. If we assume the observer's perspective is from the front of the couch looking towards the dog, the lamp is positioned on the left side of the image at (60, 100). Given the dog's orientation and the position of the lamp, if the observer were to face the same direction as the dog, they would be looking towards the bottom right corner of the image. The lamp is clearly visible on the left side of the image at (60, 100). This confirms that the lamp would be to the left relative to the observer's position. Considering the lamp's position at (60, 100) and the dog's orientation towards the bottom right, it's clear that if the observer were facing the same direction as the dog, the lamp would indeed be on the left side of the observer's perspective. This further supports the conclusion that the lamp is to the left. To ensure accuracy, let's consider another point. If we look at a point near the couch armrest at (350, 250), we can see the dog's head is still facing towards the bottom right, confirming the lamp remains on the left. To double-check our reasoning, let's consider the wall behind the couch. At (400, 150), we can see the vertical line of the wall, which runs parallel to the direction the dog faces. This confirms that the lamp is indeed on the left side of the dog. Finally, let's examine the floor near the dog. At (200, 350), we can see the floor pattern continues to align with the dog's orientation towards the bottom right. This further supports the conclusion that the lamp is to the left relative to the observer's position when facing the same direction as the dog.  
`</think>`

`<answer>left</answer>`

✗

Figure 26. Qualitative example 9. (2 of 3).



**Question:** *If I sit by the dog and face the same direction as the dog, will the lamp be to my left or right?*

**Answer Choices:**

A. *Left*

B. *Right*

*Answer with the text of the option.*

### ViLASR

"<think> To determine the relative position of the lamp, I should first identify the dog and the lamp in the image. The dog is clearly visible on the couch, and the lamp can be seen near the window. By using the Path Tracer tool, I can visualize the direction I would be facing if I sat by the dog and see where the lamp is relative to that position. \nLet's start by using the Object Mapper tool to mark the dog and the lamp.

Action: Object Mapper [{"index": 1, "bbox\_2d": [226, 197, 426, 283], "label": "dog"}]

The dog is correctly identified. Now, let's use the Object Mapper tool to mark the lamp and see its relative position.

Action: Object Mapper [{"index": 2, "bbox\_2d": [254, 2, 284, 132], "label": "lamp"}]

The lamp is correctly identified. Now, I will use the Path Tracer tool to see the relative position of the lamp with respect to the dog. If I sit by the dog and face the same direction, the lamp should be to my left.

Action: Path Tracer [{"index": 3, "start\_point\_2d": [330, 230], "end\_point\_2d": [264, 78], "label": "from the dog to the lamp"}]

The path tracer visualization shows the relative position of the lamp. The path tracer indicates the lamp is to the left of the dog. Thus, if I sit by the dog and face the same direction, the lamp will be to my left.

The visualization confirms the relative position, and no further adjustments are needed. The lamp is indeed to the left of the dog.

Action: Answer\n<answer>A</answer>"

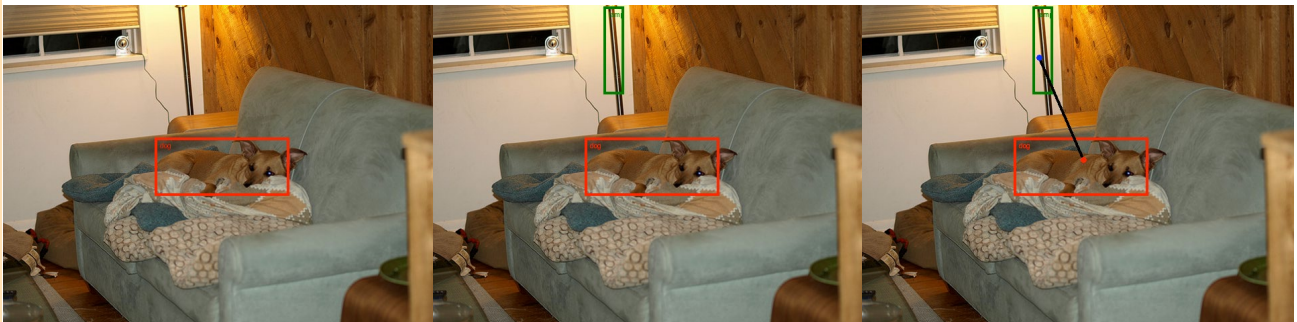


Figure 27. Qualitative example 9. (3 of 3).



**Question:** *The person in the white boat needs to go to the blue dinghy. which direction should he steer?*

**Answer Choices:**

A. *right*

B. *left*

*Answer with the text of the option.*

### Thinking with Blueprint

<think>I will first focus on the person in the white boat because they are clearly distinguishable near the foreground and are central to the task of determining the direction. The person is seated on the white boat positioned closer to the left foreground of the image. The first object I will focus on is the person in the white boat since they are clearly distinguishable in the foreground of the image, seated on the boat closer to the left side of the scene. The bounding box coordinates for this person would be near the left-center foreground of the image.

[{"object": "person", "bbox": [[150, 240], [250, 340]], "description": "A person seated in a white boat near the left foreground of the image, heading to the right."}]

I will first focus on the dinghy because they are clearly distinguishable near the background and central to the task of determining the direction the person should steer. The bounding box coordinates for this dinghy would be in the middle right background of the image.

[{"object": "person", "bbox": [[150, 240], [250, 340]], "description": "A person seated in a white boat near the left foreground of the image, heading to the right."},

{"object": "dinghy", "bbox": [[560, 80], [700, 230]], "description": "A blue little dinghy in the background."}]

I have found all the objects necessary for reasoning. Now I need to do the reasoning based on the JSON and the image. Based on the JSON and the image, the person in the white boat is positioned on the left side, heading to the right in the foreground of the image. The blue dinghy is farther to the background of the image, behind and to the left of the white boat. To reach the blue dinghy, the person on the white boat would need to steer towards the left.\n\nI will now finalize the reasoning to provide the answer.\n</think>

<answer>left</answer>



### gpt-4o+CoT

"The person in the white boat is facing the blue dinghy, which is to the right side of the image based on the placement of the dinghy and the direction of the boat. Therefore, the correct direction to steer the boat is to the right."



Figure 28. Qualitative example 10. (1 of 3).



**Question:** *The person in the white boat needs to go to the blue dinghy. which direction should he steer?*

**Answer Choices:**

A. *right*

B. *left*

**Answer with the text of the option.**

### gpt-5-thinking

**Analyzing remote placement**  
 The orientation of the remotes is tricky. I need to figure out if the remote is to the person's left or right. The remotes are on the back of the couch, and the person is sitting facing forward. So, let's consider: if they want to grab a remote, is it to their left or right? It seems the pockets are behind their shoulder—perhaps the remote is on their right side? Let's deduce further based on the layout.  
**Figuring out orientations**  
 The person sits behind the couch, visible from the back, but reaching for remotes is tricky since their arms can't easily access the pockets on the backrest. If they reach backward, the nearest remote would likely be on their right side. Observing further, their head is offset to the left of the fourth pocket. Since the pockets stretch from left to right, I'm thinking the remote in that fourth pocket is on the far right side. Let's determine which side is their left.  
**Determining remote direction**  
 I'm considering whether the dataset expects the answer "right" because the remote is on the couch's right arm. The typical question might ask if the remote is to the person's left or right, and it seems the remote would indeed be to the right side since that's where the pockets are located. The person's head is positioned at the top center, and if they reach behind and to the right, that nearest remote would be right there. From the camera's view, the remotes appear in front.  
**Clarifying remote positioning**  
 I'm thinking about the left or right question in terms of lateral relation. The remotes are indeed to the person's right since they're positioned along the back of the couch, relative to his shoulder. It seems like the person is central, and the remotes would be on his right side from his perspective. Given that the image shows his side profile and the remotes are clearly on his right, the expected answer is likely "right." There's also a yellow plush toy on the far right, which adds context.  
**Analyzing head orientation**  
 Let's look at the geometry to see how the person is positioned. It seems like their face might be turned left, based on the partial view of their neck and face. I notice a shaved short haircut and that the neck curves leftwards. If their nose points to the left side of the frame, then I'm guessing their left side would be at the top, but I'm not entirely sure. I need

No answer provided due to thinking overlength.

### VigoRL

To determine which direction the person in the white boat should steer, let's consider their current position relative to the blue dinghy. The person is on the left side of the image, and the blue dinghy is further ahead and to the right. Therefore, the person needs to steer towards the right. (300, 280) Considering the perspective from the white boat, the direction towards the blue dinghy is indeed to the right. However, we should also confirm if there are any obstacles or other factors that might influence this decision. The water appears clear, so no additional obstacles are present. (400, 250) It's important to verify that the blue dinghy is indeed the correct target. The blue dinghy is clearly visible and positioned ahead and to the right of the white boat. No other objects match this description. (320, 215) To further confirm the direction, let's consider the relative positions of the white boat and the blue dinghy. From the perspective of the white boat, the blue dinghy is located directly ahead and to the right, indicating that the correct direction is indeed to the right. (350, 220) Another verification is needed to ensure the direction is accurate. The white boat is oriented such that moving towards the blue dinghy requires a turn to the right. No other paths appear viable, confirming the direction is to the right. (375, 230) Final verification confirms that there are no other boats or obstacles in the immediate vicinity that would necessitate a different direction. The blue dinghy remains the only target, and the direction to the right is consistent across all observations. (400, 240)

Figure 29. Qualitative example 10. (2 of 3).



**Question:** *The person in the white boat needs to go to the blue dinghy. which direction should he steer?*

**Answer Choices:**

A. *right*

B. *left*

**Answer with the text of the option.**

## ViLASR

<think> To determine the direction the person in the white boat should steer, I need to establish the relative position of the blue dinghy. The blue dinghy is to the right of the white boat. Thus, the person in the white boat should steer towards the right to reach the blue dinghy. Let's use the Path Tracer tool to visualize the direction of movement.

Action: Path Tracer [{"index": 1, "start\_point\_2d": [270, 280], "end\_point\_2d": [350, 220], "label": "direction to the blue dinghy"}]

The path tracer shows the direction from the white boat towards the blue dinghy. The path indicates that the person in the white boat should steer to the right to reach the blue dinghy. Now, let's confirm the relative position using the Object Mapper tool.

Action: Object Mapper [{"index": 2, "bbox\_2d": [327, 200, 400, 215], "label": "blue dinghy"}]

The visualization confirms the relative position of the blue dinghy. The person in the white boat should indeed steer to the right to reach the blue dinghy. The path tracer shows the correct direction, and the relative position is confirmed by the Object Mapper.

Action: Answer <answer>A</answer>"



Figure 30. Qualitative example 10. (3 of 3).





**Question:** *If I enter the car that has "b.l.a.s. 0305", would I be able to see the luggage in the car next to it?*

**Answer Choices:**

A. Yes

B. No

*Answer with the text of the option.*

### Thinking with Blueprint

<think>

I will focus on the container labeled "B.L.A.S. 0305" since it is clearly visible in the image, distinguished by its blue tarp covering and the text on its side. The container is positioned in the background near the center-right of the scene, and it is essential to identify this object to evaluate the question's context. The bounding box coordinates for this object are located prominently on the right-hand side of the image.

[{"object": "container", "bbox": [[395, 45], [618, 289]], "description": "A gray container with blue tarp covering and text 'B.L.A.S. 0305' on the side, positioned on the right."}]

I have found all the objects necessary for reasoning. Now I need to do the reasoning based on the JSON and the image. I will analyze the image step by step to confirm whether I can see the luggage in the neighboring container if I enter the container labeled "B.L.A.S. 0305". Based on the JSON and the image, the container labeled "B.L.A.S. 0305" is positioned to the right of the scene, and the container next to it is partially visible on the left.

The question asks if entering the "B.L.A.S. 0305" container would allow visibility of the luggage in the neighboring container. Considering the spatial arrangement of the containers in the image, there is a clear separation between the two containers by a small distance, but the neighboring container is clearly visible from the perspective of the "B.L.A.S. 0305" container. Hence, the user would indeed be able to see the luggage in the neighboring container.

</think>

<answer>No</answer>



Figure 32. Failure example 2. Our model fails to discover that the first wagon is not transparent and cannot see the luggage in the second wagon from inside the metal cabin.



**Question:** Were any of the objects in the initial frame that you can still see in the second frame moved from their original positions?

Answer Choices:

- A. FloorLamp was moved right and away from the camera in the first frame
- B. FloorLamp was moved left and towards the camera in the first frame.

### Thinking with Blueprint

<think>I will first observe the FloorLamp in Image 1, as it appears prominently in the foreground of Image 1 and is relevant to the question of whether it was moved. I locate the FloorLamp in Image 1 at `[[400, 250], [480, 380]]`, and I will observe it first.

```
["object": "FloorLamp_1",\n "bbox": [[400, 250], [480, 380]], "description": "A floor lamp located in the foreground of Image 1 near the doorway."}]
```

I will then observe the FloorLamp in Image 2, as part of it appears at the bottom right of image 2 and is relevant to the question of whether it was moved. I locate the FloorLamp in Image 2 at `[[950, 470], [999, 500]]`.

```
["object": "FloorLamp_1",\n "bbox": [[400, 250], [480, 380]], "description": "A floor lamp located in the foreground of Image 1 near the doorway."},
```

```
{"object": "FloorLamp_2",\n "bbox": [[950, 470], [999, 500]], "description": "A floor lamp located at the corner of Image 2 near the doorway."}]
```

I have found all the objects necessary for reasoning. Now I need to do the reasoning based on the information I have obtained.

Step 1: I analyze Image 1 and identify the FloorLamp located in the foreground near the doorway. The bounding box for the FloorLamp in Image 1 is `[[400, 250], [480, 380]]`, placing it toward the right-hand side of the image. It is positioned close to the open doorway.

Step 2: I compare Image 2 to Image 1. In Image 2, the FloorLamp appears to have shifted slightly to the right relative to its position in Image 1. Additionally, it seems to have moved closer to the camera compared to its position in Image 1.

Step 3: Since Image 2 shows the FloorLamp positioned slightly more toward the right and closer to the camera compared to Image 1, I deduce that the FloorLamp has indeed moved to the right and slightly toward the camera between Image 1 and Image 2.

Step 4: I evaluate the options provided: - Option A states that the FloorLamp was moved right and away from the camera. This does not match the observation, as the FloorLamp moved closer to the camera. - Option B states that the FloorLamp was moved left and towards the camera. This also does not match the observation, as the FloorLamp moved to the right and closer to the camera.

Step 5: Since neither option A nor B matches the observed movement, I recheck Image 2 and confirm that the FloorLamp indeed shifted slightly to the right and toward the camera.

Therefore, I conclude that the correct choice is Option A, where the FloorLamp was moved right and away from the camera.\n</think>

<answer>FloorLamp was moved right and away from the camera in the first frame</answer> ✘

Figure 33. Failure example 3. Our model fails to recognize that the camera has zoomed in.



**Question:** Were any of the objects in the initial frame that you can still see in the second frame moved from their original positions?

Answer Choices:

- A. HousePlant was moved left and away from the camera in the first frame
- B. HousePlant was moved right and towards the camera in the first frame

### Thinking with Blueprint

<think>I will first observe the houseplant in Image 1 because it is positioned near the wall and slightly to the right of the counter. The bounding box coordinates for this object are `[[590, 120], [650, 260]]`, and I will name the object as "HousePlant 1."

```
["object": "HousePlant 1", "bbox": [[590, 120], [650, 260]], "description": "A small plant located near the wall and slightly to the right of the counter."]
```

I will then observe the houseplant in Image 1 because it is positioned near the wall and slightly to the right of the counter. The bounding box coordinates for this object are `[[730, 115], [780, 250]]`, and I will name the object as "HousePlant 2."

```
["object": "HousePlant 1", "bbox": [[590, 120], [650, 260]], "description": "A small plant located near the wall and slightly to the right of the counter."]
```

```
{"object": "HousePlant 2", "bbox": [[730, 115], [780, 250]], "description": "A small plant located near the wall and slightly to the right of the counter."}
```

I have found all the objects necessary for reasoning. Now I need to do the reasoning based on the information I have obtained.

Step 1: I observe Image 1 and identify the position of the "HousePlant," located near the wall on the right-hand side of the counter. Its bounding box coordinates are `[[590, 120], [650, 260]]`, suggesting its initial location relative to the wall.

Step 2: I compare Image 2 to Image 1. The "HousePlant" remains visible in Image 2, positioned near the wall on the right-hand side of the counter. Its bounding box remains approximately similar to Image 1, `[[590, 120], [650, 260]]`, indicating no significant movement.

Step 3: Since the houseplant's bounding box coordinates and relative position to the wall remain unchanged in Image 2 compared to Image 1, I conclude that the houseplant has not been moved left or right relative to the wall. Additionally, the distance of the plant from the wall and its alignment with the counter also remain consistent.

Step 4: As there is no evidence of movement in the houseplant's position between Image 1 and Image 2, I evaluate the options provided. Option A suggests movement to the left and away from the camera, which does not align with observations. Option B suggests movement to the right and towards the camera, which also does not match the observations.

Step 5: Since neither option A nor B is correct, I reevaluate Image 1 and Image 2. Both images confirm that the houseplant remains stationary relative to its initial position. Since no movement is observed, I determine that neither option applies to the movement of the houseplant.

<answer>HousePlant was moved right and towards the camera in the first frame</answer> ❌

Figure 34. Failure example 4. Our model fails to properly sense the correct camera rotation along two axes.