

—Supplementary Material—

SpatialDreamer: Incentivizing Spatial Reasoning via Active Mental Imagery

Meng Cao^{1*}, Xingyu Li^{1*}, Xue Liu¹, Ian Reid¹, Xiaodan Liang^{1,2†}

¹Mohamed bin Zayed University of Artificial Intelligence ²Sun Yat-sen University

*Authors contributed equally to this research. †Corresponding author.

To supplement the findings presented in the main paper, this supplementary material offers additional quantitative and qualitative results. The specific contents are as follows:

- Statistics of SpatialDreamer-SFT dataset.
- Visualizations of response length.
- Prompts for SpatialDreamer-SFT Curation.
- Evaluation prompts.
- Qualitative results of SpatialDreamer.

Statistics of SpatialDreamer-SFT dataset. Table 1 summarizes the key statistics of the SpatialDreamer-SFT dataset. It contains two types of samples: single-pass and reflective reasoning traces. The single-pass subset includes 1,334 annotated trajectories with an average of 2.7 reasoning steps per trace, representing direct spatial reasoning without correction. The reflective subset includes 392 trajectories with an average of 3.5 reasoning steps, where each trace may include revisions based on prior feedback. On average, 1.2 errors are intentionally injected per reflective sample to simulate self-correction behavior.

Visualizations of response length. Figure 1 compares the evolution of response lengths between our GeoPO and the vanilla GRPO across training steps. GeoPO maintains a stable response length throughout the optimization process while GRPO rapidly collapses to very short outputs (below 50 tokens) within the first 20 steps. These results highlight the effectiveness of our GeoPO in preserving expressive and content-rich responses via step-wise policy optimization.

Prompts for SpatialDreamer-SFT Curation. The prompts for single-pass samples and reflective reasoning samples in the SpatialDreamer-SFT dataset are shown in Figure 4 and Figure 5, respectively.

Evaluation Prompts. The prompts for evaluation on benchmarks of SAT [1], MindCube [3], and VSI-Bench [2] are shown in Figure 6.

Qualitative results of SpatialDreamer. Figure 2 and Figure 3 present the qualitative results of our SpatialDreamer in both simulated and real-world environments. Specifically, Figure 2a illustrates a simulated indoor scenario, where the agent reasons about how to face the window. This exam-

Table 1. **Key statistics of the SpatialDreamer-SFT dataset.** “Average reasoning steps per trace” represents the average number of interaction rounds. “Injected errors per reflective sample” measures the average number of injected errors per reflective reasoning trace.

Statistics of SpatialDreamer-SFT	Value
Single-pass samples	
Total annotated trajectories	1334
Average reasoning steps per trace	2.7
Reflective samples	
Total annotated trajectories	392
Average reasoning steps per trace	3.5
Injected errors per reflective sample	1.2

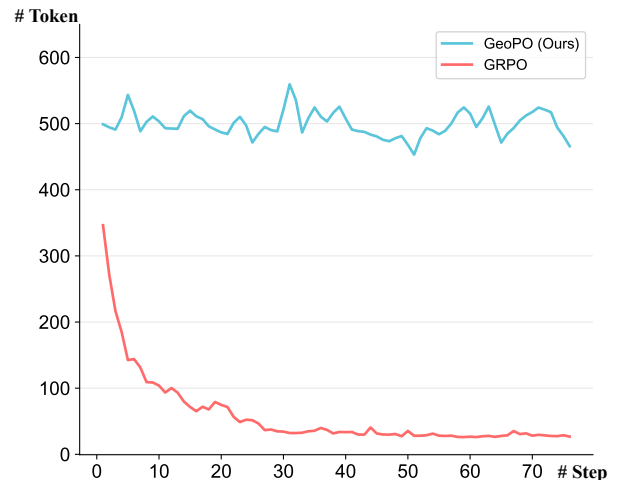


Figure 1. **Comparison of response length** during training. GeoPO maintains stable and informative responses, while GRPO collapses to short outputs.

ple explicitly demonstrates the SpatialDreamer’s ability of *reflective reasoning*, where the agent first moves forward to gather additional spatial cues, then reanalyzes the up-

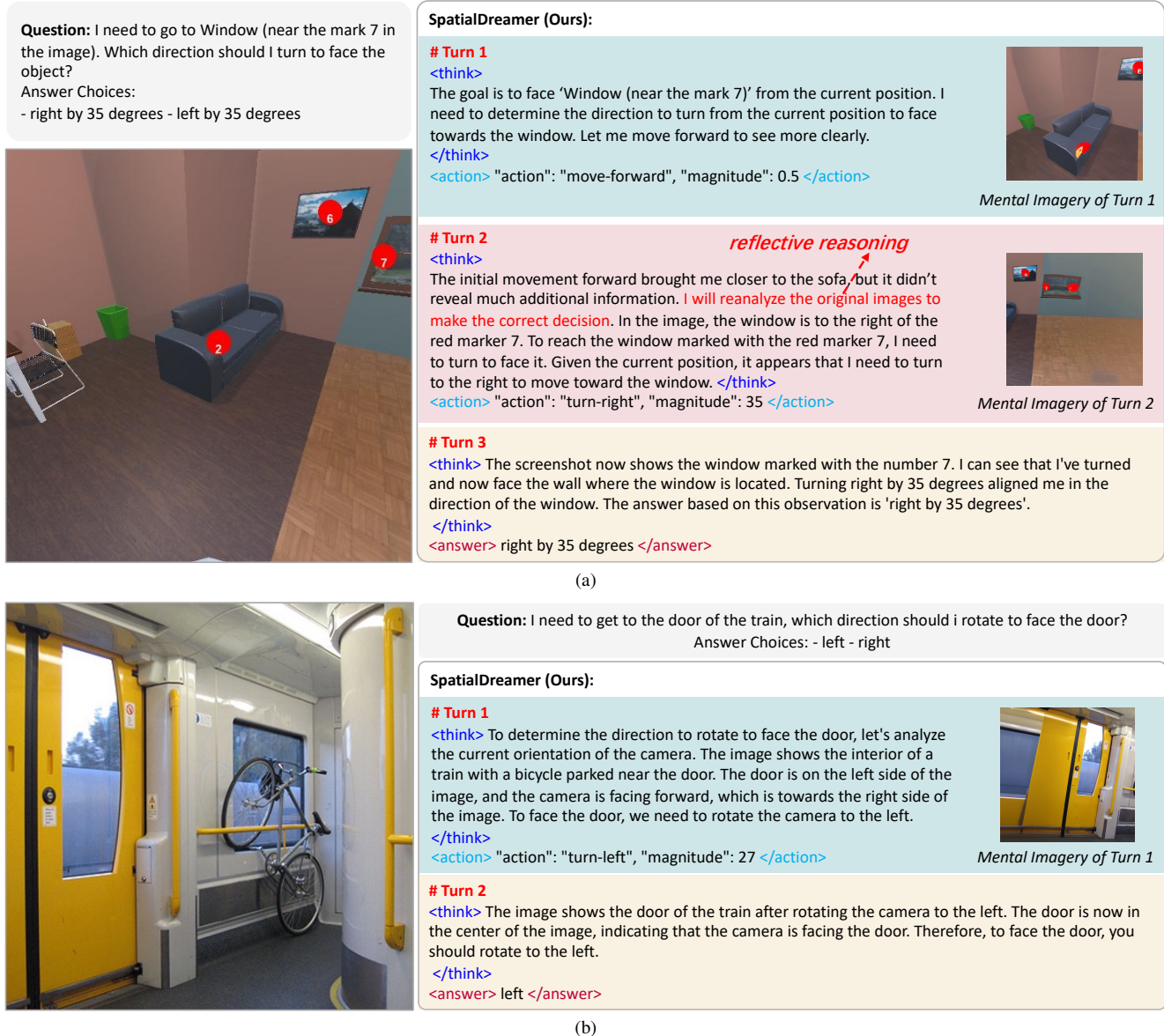


Figure 2. Qualitative results of our SpatialDreamer.

dated view before making the final directional decision. In contrast, Figure 2b shows a real-world scene inside a train, where SpatialDreamer determines the rotation direction needed to face the door. Together, these examples highlight the generalization of our SpatialDreamer from simulation to real-world settings and its capability to perform multi-step spatial reasoning when necessary.

References

[1] Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, et al. Sat: Dy-

namic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024. 1

[2] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025. 1

[3] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*, 2025. 1



Question:
Based on these four images (image 1, 2, 3, and 4) showing the red ball from different viewpoints (front, left, back, and right), with each camera aligned with room walls and partially capturing the surroundings: From the viewpoint presented in image 1, what is to the left of the red ball?
A. Wall B. Smoking machine
C. Cardboard-covered sliding door D. Window

SpatialDreamer (Ours):

Turn 1

<think>

From image 1, the red ball is on a quilt or mattress, and in front of it there is a device with a silver exhaust pipe, likely a kitchen range hood. Image 2 is taken from the red ball's left side, showing a window on that side. Image 3 is from behind the red ball, revealing a sliding door covered with cardboard. Image 4 is from the right side, showing a smooth tiled wall. Starting from image 1 as the initial viewpoint, to determine what is on the left of the red ball, I turn left to inspect the scene.

</think>

<action>

"action": "turn-left", "magnitude": 45

</action>



Mental Imagery of Turn 1

Turn 2

<think>

After turning left, the wall becomes visible, indicating that from the image 1 viewpoint, the left side of the red ball corresponds to the wall.

</think>

<answer> A. Wall </answer>

(a)



Question:

Based on these four images (image 1, 2, 3, and 4) showing the small white table and plastic container from different viewpoints (front, left, back, and right), with each camera aligned with room walls and partially capturing the surroundings: From the viewpoint presented in image 2, what is to the back of the small white table and plastic container?

A. TV B. Leather loveseat with three seat cushions C. Leather loveseat D. Two single sofas

SpatialDreamer (Ours):

Turn 1

<think>

Image 2 serves as the reference viewpoint, showing the small white table and plastic container in the middle of the room, with a patterned cushion on top. Image 1 shows the front view, facing the TV side, with two sofas visible (one on the left and one on the right). Image 3 shows the left side of the object, where two single sofas are visible along with a chair and small table. Image 4 shows the back view, where the TV and entertainment stand are clearly visible. Thus, from the position and direction of image 2, we need to turn right to determine what is to the back of the small white table and plastic container.

</think>

<action> "action": "turn-right", "magnitude": 90 </action>



Mental Imagery of Turn 1

Turn 2

<think>

After turning right, the TV comes into view. Based on the relative position, it can be determined that from the position and direction of image 2, the TV is located behind.

</think>

<answer> A. TV </answer>

(b)

Figure 3. Qualitative results of our SpatialDreamer.

Prompts for SpatialDreamer-SFT (single-pass)

Prompt for tool call

system_prompt = """"You are an expert spatial reasoning assistant. Given a spatial reasoning task, suggest ONE tool call that would help verify or explore the spatial relationships.

Available tools:

- generate_svc_image: Generate images by simulating camera movements
- action: "move-forward" (magnitude in meters, e.g., 0.75-2.0)
- action: "turn-left" (magnitude in degrees, e.g., 27-90)
- action: "turn-right" (magnitude in degrees, e.g., 27-90)
- image_index: which image to use as starting point (0 for first original image, or use the index of a previously generated image)

Note: You will see both original images and images generated by previous tool calls. You can use any of them as the starting point.

Respond in JSON format:

```
{
  "action": "turn-right",
  "magnitude": 27,
  "image_index": 0,
  "reasoning": "Turn right from image 1 to verify the object's position on the right side"
}"""
```

first tool call

user_prompt = f"""

Task Context: {task}

Initial Reasoning: {init_reason}

You have {len(image_paths)} original image(s) available.

Please suggest the FIRST tool call that would help explore the spatial relationships.

"""

second tool call

user_prompt = f"""

Task Context: {task}

Initial Reasoning: {init_reason}

Previous tool execution history:

Step {idx}:

- Reasoning: {prev_step.get('reasoning', 'N/A')}\n"
- Tool Call: {json.dumps(prev_step.get('tool_call', {}), ensure_ascii=False)}\n"
- Tool Response: {prev_step.get('tool_response', 'N/A')} <image>\n"

You now have {len(image_paths)} image(s) available "

(including original images and images generated by previous tool calls).

You can use any of them as the starting point for the next tool call.

Please suggest the SECOND tool call based on the previous observations and available images.

"""

Prompt for answer generation

system_prompt = """"You are an expert spatial analyst. Based on the task, initial reasoning, and tool execution results, provide a final comprehensive reasoning and answer. You will see both original images and images generated by tool calls. Use all available visual information to make your final assessment."""

user_prompt = f"""Task Context: {human_prompt}

Initial Reasoning: {think_text}

Tool Execution History: {chr(10).join(tool_history)}{image_info}

Please synthesize all the information above (including all available images) and provide your final reasoning and answer. Format your response strictly as: <think>Your comprehensive reasoning here</think><answer>Your final answer here</answer>"""

Figure 4. Prompts for generating single-pass reasoning samples in SpatialDreamer-SFT dataset.

Prompts for SpatialDreamer-SFT (Reflective Reasoning)

Prompt for tool call

system_prompt = """You are an expert spatial reasoning assistant.

Available tools:

- generate_svc_image: Generate images by simulating camera movements
- action: "move-forward" (magnitude in meters, e.g., 0.75-2.0)
- action: "turn-left" (magnitude in degrees, e.g., 27-90)
- action: "turn-right" (magnitude in degrees, e.g., 27-90)
- image_index: which image to use as starting point (0 for first original image, or use the index of a previously generated image)

Note: You will see both original images and images generated by previous tool calls. You can use any of them as the starting point.

Respond in JSON format:

```
{  
  "action": "turn-right",  
  "magnitude": 27,  
  "image_index": 0,  
  "reasoning": "Turn right from image 1 to verify the object's position on the right side"  
}
```

first tool call

user_prompt = f"""

Task Context: {task}

Initial Reasoning: {init_reason}

You have {len(image_paths)} original image(s) available.

Important constraint for this step:

Deliberately choose an action that is unlikely to help answer the task (i.e., an uninformative or irrelevant camera move), while still being a valid tool call.

Please suggest the FIRST tool call. It should:

Be valid syntactically.

Be spatially plausible but unhelpful for clarifying the task.

Include concise reasoning explaining why this action is (ironically) not helpful.

"""

second tool call

user_prompt = f"""

Task Context: {task}

Initial Reasoning: {init_reason}

Previous tool execution history:

Step {idx}:

- Reasoning: {prev_step.get('reasoning', 'N/A')}\n"
- Tool Call: {json.dumps(prev_step.get('tool_call', {}), ensure_ascii=False)}\n"
- Tool Response: {prev_step.get('tool_response', 'N/A')} <image>\n"

You now have {len(image_paths)} image(s) available "

(including original images and images generated by previous tool calls).

You can use any of them as the starting point for the next tool call.

Your objective for this step:

Recognize that the first action was unhelpful.

Provide a corrective, task-relevant action that maximizes information gain about the spatial relationships in question.

Justify how this new action addresses the deficiency of the first action and leverages available images.

Please suggest the SECOND tool call.

"""

Prompt for answer generation

system_prompt = """You are an expert spatial analyst. Based on the task, initial reasoning, and tool execution results, provide a final comprehensive reasoning and answer. You will see both original images and images generated by tool calls. Use all available visual information to make your final assessment."""

user_prompt = f"""Task Context: {human_prompt}

Initial Reasoning: {think_text}

Tool Execution History: {chr(10).join(tool_history)}{image_info}

Please synthesize all the information above (including all available images) and provide your final reasoning and answer.

Format your response strictly as: <think>Your comprehensive reasoning here</think><answer>Your final answer here</answer>"""

Figure 5. Prompts for generating reflective reasoning samples in SpatialDreamer-SFT dataset.

Evaluation Prompt

```
f""""<|im_start|>system
You are a helpful assistant.

# Tools

You may call one or more functions to assist with the user query.

You are provided with function signatures within <tools></tools> XML tags:
<tools>
{"type": "function", "function": {"name": "image_generator", "description": "A tool for generating image by taking actions in
3D space using the SVC (Stable Virtual Camera) world model.", "parameters": {"type": "object", "properties": {"action":
{"type": "string", "description": "The action to take in the 3D space. Must be one of: 'move-forward' (move camera forward),
'turn-left' (rotate camera left), 'turn-right' (rotate camera right)", "enum": ["move-forward", "turn-left", "turn-right"]},
"magnitude": {"type": "number", "description": "The magnitude of the action. For 'move-forward', this is the distance in
meters (e.g., 0.75). For 'turn-left' or 'turn-right', this is the angle in degrees (e.g., 27)", "image_index": {"type": "integer",
"description": "The index of the image to generate. The first image is 0."}}, "required": ["action", "magnitude",
"image_index"]}}}}
</tools>

For each function call, return a json object with function name and arguments within <tool_call></tool_call> XML tags:
<tool_call>
{"name": <function-name>, "arguments": <args-json-object>}
</tool_call><|im_end|>
<|im_start|>user
Answer the given question. You must conduct reasoning inside <think> and </think> first every time you get new information.
After reasoning, if you find you lack some knowledge, you can call the image generator by <tool_call> query </tool_call> to
explore the 3D space by taking actions like moving forward or turning and it will return the generated image based on your
selected image and action between <tool_response> and </tool_response>. You can call the tool as many times as your want.
If you find no further external knowledge needed, you can directly provide the answer inside <answer> and </answer>,
without detailed illustrations. For example, <answer> Beijing </answer>. Question: {Task}
<|im_end|>
<|im_start|>assistant
""""
```

Figure 6. Prompts for evaluation on SAT, MindCube, and VSI-Bench.