

# Abstract 3D Perception for Spatial Intelligence in Vision-Language Models

## Supplementary Material

### 6. Implementation Details

#### 6.1. Hyperparameters

We implement SandboxVLM using GPT-5-mini as the backbone VLM, and for variants, we also test with GPT-4o and GPT-5. The results are from the OpenAI API as of September 2025 to November 2025. For abstract control, we allow the VLM to choose from: Left, Forward-left, Forward, Forward-right and Right, which are later converted to trajectories among (1) turning left or right 90° (2) moving forward 0.5m then turning left or right 36° (3) moving forward 1.0m. For multi-view priors, we use SEVA [41] to generate 4 additional views per selected trajectory. We utilize the pointmap from VGGT [29] as the 2D-to-3D lifting model, and the number of views for Multi-View Voting  $N$  is set to 3 to balance accuracy and point numbers. In our implementation,  $M = 3$ ,  $T = 8$ ,  $N = 2$ ,  $\delta = 0.5m$ .

#### 6.2. 3D Reasoning Prompt

We perform 3D reasoning on the rendered images from 3D Sandbox. The prompt is shown as follows:

Listing 1. System prompt for spatial reasoning tasks.

```
sys_prompt = (
    "You are an AI assistant designed to help
      with spatial reasoning in a 3D indoor
      scene. "
    "You must analyze any provided images or
      observations and answer the question.\n\n"
    "Rules:\n"
    "1. You should output the exact answer from
      the choices.\n"
    "2. You will be provided with multiple
      imagined views if you taking
      corresponding actions to help you answer
      the questions.\n"
    "3. Your response should be formatted like
      <think>your analysis for the question and
      the images, think step by
      step.</think><answer>your exact answer
      selected from the choices, do not include
      other words</answer> Your score will
      depend on the answer, but thinking
      clearly and make use of the clues and
      common sense can help you improve your
      answer.\n"
```

This simple prompt design guides the VLM to systematically perceive and reason about the provided question and 3D context from 3D Sandbox.

#### 6.3. Subsampling for SAT-Synth

Since SAT-Synth has 4000 data samples, to control the cost with the OpenAI api, we subsample 98 examples from it, an

amount comparable with the 150 samples in SAT-Real split. The subsampled items are listed as follows:

Listing 2. The indices of subsampled tasks from SAT-Synth.

```
val: 1023 1177 13 1407 1727 1932 2057 2279 248
     2790 299 3170 33 3437 3790 3952 480 571 657
     788 1039 1218 1321 1418 1793 1934 2062 2327
     2510 2825 3027 3181 3324 3652 3827 3964 481
     619 662 817 1054 1244 1351 1523 1885 2001
     2114 2433 2541 2826 3041 3184 3336 3667 3873
     3978 509 623 667 93 1076 1261 1357 1559 190
     2027 214 2466 259 283 3051 3241 3362 3707
     3939 41 514 635 691 results.json 113 1272
     1397 1717 1900 2035 2157 2476 2666 2954 3152
     3257 3381 3718 3947 416 544 639 761
```

#### 6.4. Details of Benchmarks

We present the examples of the benchmarks involved in our experiments in Table 4, for readers' information.

#### 6.5. Ablation Design

During all our ablation studies listed in 4.3, we observe the prompting method presented in Sec. 3.5 and Sec. 6.2 to compare the representation side of the design fairly. We further detail the composition of each ablation study experiment in Table 5.

### 7. Error Analysis

Error propagation is inherent to modular pipelines. In our framework, error sources include video diffusion, VLM pointing, image segmentation, and depth/camera estimation. We provide qualitative visualizations and human-verified statistics to analyze these errors.

Most failures stem from inaccurate VLM pointing, which often propagates to or reinforces image segmentation errors. Imprecision in video generation and depth or camera estimation primarily introduces dispersed 3D point noise. Our abstraction and the Multi-View Voting mechanism (Section 3.4) are explicitly designed to suppress the propagation of these errors and imprecisions.

### 8. More Analysis

#### 8.1. SAT Benchmark

Across different backbones on SAT-Real, we observe consistent gains in the Ego-Movement, Goal-Aiming, Action-Consequence, and Perspective subtasks, with particularly strong results on Ego-Movement (exceeding 90% accuracy and reaching 100.0% in some settings). We attribute these improvements to the explicit 3D spatial representation and

Table 4. Question types and examples from each benchmark. The four types of benchmarks we evaluated on contain various tasks from metric questions like relative depths to comprehensive ones like physical scene understanding. For more details, please refer to the corresponding paper.

<i>SAT Bench</i> [25]	
Ego Movement	How did the camera likely rotate when shooting the video?
Object Movement	Were any of the objects in the initial frame that you can still see in the second frame moved from their original positions?
Goal Aim	I need to go to the houseplant. Which direction should I rotate to face it?
Action Consequence	I need to go to the mirror (marked 1). If I turn left by 90 degrees, will I be facing away from the mirror?
Perspective	For someone at the x marked point and facing 90 degrees to the left, will the person surfing be to their left or right?
<i>BLINK</i> [12]	
Relative Depth	Which point is closer to the camera?
Spatial Relation	Is the bed at the right side of the dining table?
<i>EmbSpatial Bench</i> [9]	
Spatial Relation	How are statue and tissuebox positioned in relation to each other in the image? (multiple choices)
	A. The statue is at the left side of the tissuebox.
	B. The statue is touching the tissuebox.
	C. The statue is outside the tissuebox.
	D. The statue is right of the tissuebox.
<i>PhysBench</i> [6]	
Physical Object Property	Which color of balls has the largest number?
	Given that the applied force is the same, which object in the images has higher stiffness?
Physical Object Relationships	What is the distance between the yellow cube and the blue ball? (The blue cube has a width of 2 cm.)
	Which car has a higher average speed?
Physical Scene Understanding	How does the focal length of the camera change?
	What causes the change in water level in the cup?
Physical-based Dynamics	Which can is the ball most likely to land in?
	What is the correct sequence of images to make a gift box containing the perfume bottle?

Table 5. Composition of each ablation method. We list: Average: the average performance on SAT-Real [25]; Modality: in which modality is the context information provided to VLM; Priors: whether a world model is used as a 3D-aware prior in generating the context; 3D: whether the context is lifted into the 3D space; Filtered: before constructing the context, whether a VLM is used to collect information relevant with the task; Using Boxes: whether we finally construct the 3D Sandbox as the context.

No.	Method	Average	Modality	Priors	3D	Filtered	Using Boxes
(1)	Vanilla VLM	75.4	-	No	No	No	No
(2)	Scene-Graph Text Prompt	77.0	Text	Yes	No	No	No
(3)	Multi-View Images Only	78.7	Vision	Yes	No	No	No
(4)	Rendered Point Clouds	73.7	Vision	Yes	Yes	No	No
(5)	3D Coordinate Text Prompt	<u>80.8</u>	Text	Yes	Yes	Yes	Yes
(6)	Rendered Proxy Points	77.0	Vision	Yes	Yes	Yes	No
(7)	Full SandboxVLM	<b>84.1</b>	Vision	Yes	Yes	Yes	Yes

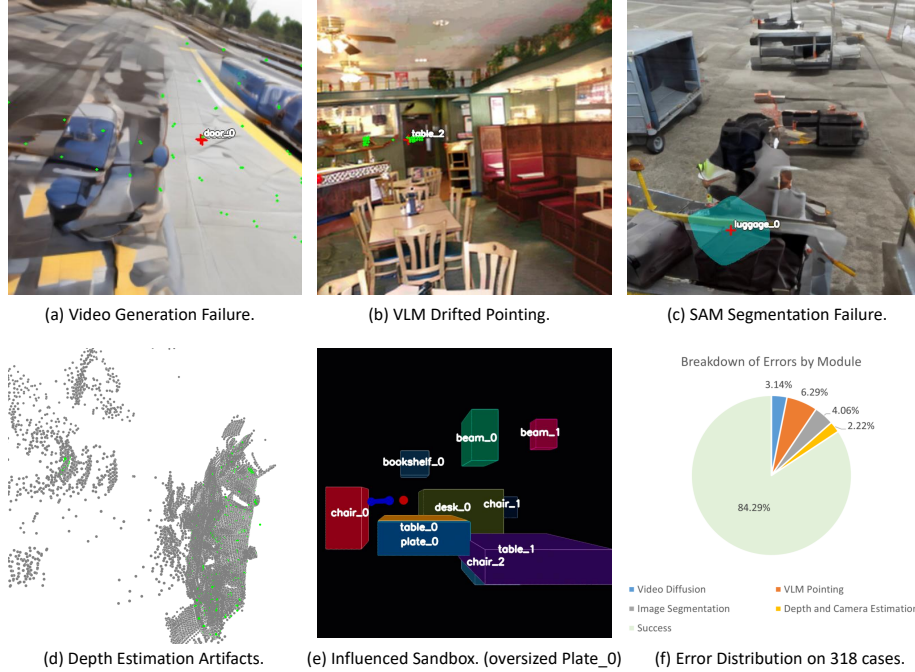


Figure 5. (a)-(e): Different modes of failure cases. (f) Error distribution statistics on SAT-Real. Please zoom in for a better view.

reasoning process in our framework, which effectively captures view changes within a stable and coherent 3D context. For Perspective tasks—where the model must reason from another agent’s or object’s viewpoint—SandboxVLM also achieves notable gains, benefiting from its ability to reconstruct only relevant regions of the 3D scene and represent objects via 3D-aware bounding boxes.

The relatively lower performance on the Object-Movement subtask compared with MindJourney, GPT-4o as backbone on SAT-Synth, likely stems from the static nature of our current 3D reconstruction module. Extending SandboxVLM toward dynamic 3D reconstruction would thus be a promising direction for future research.

## 8.2. Representation Analysis

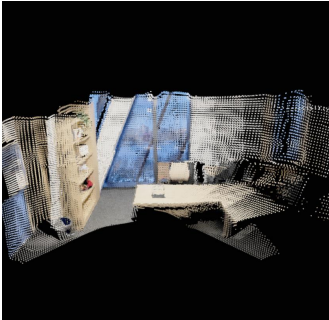
We provide analysis of failure cases from different types of inputs in Fig. 6.

**(a) Rendered Point Clouds.** The sparse and incomplete nature of the point cloud representation can lead to misinterpretation of spatial relationships. In the example, the VLM misjudges that the bookshelf also appears on the right side, causing confusions in interpreting the spatial layout. This flaw highlights the limitations of relying solely on point cloud rendering for spatial reasoning tasks, pointing to a gap between the 3D reconstruction research and VLMs, a gap which our work aims to bridge.

**(b) Scene-graph Text Prompt.** The 3D coordinate text prompt provides filtered and structured information about object locations in the text format. However, in most cases

we observed that the VLM began reasoning in pure text format with the provided numerical coordinates. It even started computing vectors and cross products to deduce spatial relationships, without attending to visual signals. We attribute this to the strong bias towards the language modality in VLM CoT, where the model tends to prioritize textual reasoning over visual cues when both are present. This also validates our methods’ motivation in constructing a visual 3D context for the VLM to perceive.

**(c) Ours.** We also analyze two typical failure modes in our framework: lack of appearance details, and lack of modeling dynamics. In (c.1), we demonstrate a case where the question involves the original facing direction of the human inside the scene. However, our 3D Sandbox does not model the exact initial facing direction of the human, leading to hallucinations. This is due to our focus on geometric cues rather than appearance in the representation. Future work addressing appearance modeling in 3D Sandbox, e.g. with Sam-3D-Objects, would help mitigate this issue. In (c.2), the VLM fails to correctly discern the object movement between two 3D Sandbox renderings. We attribute this to (i) the errors introduced in the reconstruction pipeline such as VLM pointing and the 3D reconstruction from generated views, which can misalign objects slightly; (ii) the static nature of our current 3D reconstruction module, which cannot model dynamic changes directly in one 3D Sandbox. Addressing these limitations would be a promising direction for future research.

(a)  **How did the camera likely rotate when shooting the video?**

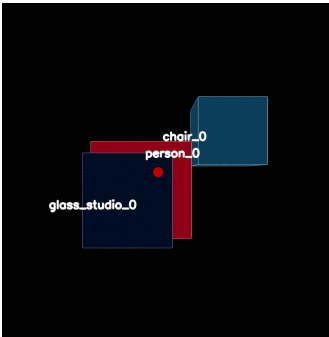
<think>The point-cloud shows the **bookshelf and desk repeated/extended toward the right side** in the second image, indicating the scanner swept that direction during capture. The provided imagined views for turning left move the shelf toward center/left, which is the opposite of the sweep seen in the point cloud. Therefore the camera most likely rotated to the right while shooting. </think> <answer> rotated right</answer>

**For someone near the white table facing towards the left, will the windows be to their left or right?**

(b) <think>Person is near the white table (table center = [0.384,0.263,0.759]). Window center = [0.289,-0.125,1.416]. Facing "left" means facing negative X (forward = (-1,0,0)). **Compute vector from person to window:** window - table = [-0.095, -0.388, 0.657]. The person's right vector when facing negative X is forward x up = (-1,0,0) x (0,1,0) = (0,0,-1). Dot product of relative vector with right =  $[-0.095, -0.388, 0.657] \cdot [0,0,-1] = -0.657$  (negative), which means the window lies to the person's left. </think> <answer> left</answer>

**3D Scene Object Locations:**

```
table_0 (table): \n Center: [0.384, 0.263, 0.759] \n Size [0.148, 0.055, 0.196] \n Min corner: [0.310, 0.235, 0.661] \n Max corner: [0.458, 0.290, 0.857] \n Distance from origin: 0.890m \n \n window_0 (window): \n Center: [0.289, -0.125, 1.416] \n Size [0.196, 0.311, 0.121] \n Min corner: [0.191, -0.280, 1.356] \n Max corner: [0.387, 0.031, 1.477] \n
```

(c.1)  **The person sitting on the chair needs to face the glass studio. Which direction should the person sitting on the chair turn to face the glass studio?**

<think>The bird's-eye render shows the glass\_studio\_0 located to the person's left-front relative to **their current facing direction**. To point toward the studio the person must turn toward their left rather than right. The needed turn is a moderate angle (about 45 degrees) left to face the glass studio. </think> <answer> rotate left by 45 degrees</answer>

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

(c.2) <think>I compare the initial egocentric image and the helper (second) egocentric image: the same desk, monitors, cup, and a small cardboard box near the trash bin are visible in both frames. The rendered bird's-eye and step-back views for t0 and t1 show **the same object clusters with no relative translations** that would indicate an object was moved. The imagined spin views of the corner with the cardboard box also show the box in the same position relative to the bin and desk. There is no evidence any visible object (including the cardboard box) was shifted left or away from the camera between frames. Therefore no object moved. </think> <answer> no object moved</answer>

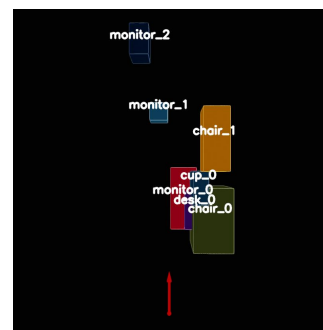


Figure 6. **Examples of failure cases.** We show failure cases from different types of inputs: (a) Rendered Point Clouds; (b) 3D Coordinate Text Prompt; (c) Ours. The erroneous messages are highlighted in red. We provide analysis of these failure modes in Sec. 8.2.