

Appendices

In this supplementary material, we present additional details and clarifications that are omitted in the main text due to space constraints.

- [Appendix A](#) Limitations.
- [Appendix B](#) Dataset Details.
- [Appendix C](#) Implementation Details.
- [Appendix D](#) More Results.

A. Limitations

While ROBOSPATIAL significantly improves spatial reasoning capabilities in VLMs, certain design choices naturally introduce trade-offs and areas for future exploration.

First, the dataset relies on a top-down occupancy map to identify and annotate empty regions for spatial context and compatibility tasks. This approach simplifies reasoning about object placement on horizontal surfaces and enables efficient data generation, but it currently does not support spatial questions involving containment—such as whether an object can fit inside or under another object—which would require more detailed volumetric modeling.

Second, although the models are deployed on a real robot using a modular approach, we do not yet explore tighter forms of integration such as training it jointly with robot trajectories [27]. Investigating these alternatives could enhance downstream policy learning and enable more seamless end-to-end systems.

Finally, ROBOSPATIAL focuses on indoor and tabletop scenes containing objects commonly encountered in household environments, and does not include humans or animals. This reflects the nature of source datasets and our emphasis on robot object manipulation. While this limits coverage of social or dynamic interaction scenarios, trained models still generalize well to out-of-distribution benchmarks like BLINK, which include humans and animals—suggesting that the learned spatial representations are broadly transferable.

B. Dataset Details

B.1. Dataset Statistics

We provide the full dataset statistics in Tab. 7. For all training, we use only 900,000 spatial relationships, sampled equally across all datasets, due to computational constraints. We further experiment on the effect of data scaling on Tab. 9 and explain the results. Notably, HOPE [54] and GraspNet-1B [12] contain similar tabletop images captured from different perspectives, resulting in lower dataset diversity for the tabletop environment. We plan to enhance the diversity of ROBOSPATIAL by incorporating additional tabletop datasets.

B.2. Choice of Spatial Relationships

In designing the dataset, we focused on spatial relationships that directly impact robotic perception, planning, and interaction: context, compatibility, and configuration. These were selected to reflect the core spatial reasoning challenges that robots encounter when operating in complex, real-world environments.

We intentionally excluded tasks such as object counting, as we consider them to fall outside the scope of spatial understanding. While counting is an important visual reasoning skill, it does not require reasoning about spatial relations between objects or between objects and their environment. For example, determining that “three cups are on the table” is a perceptual task rather than a spatial reasoning one. As such, counting may complement but does not substitute for the types of relational reasoning we target. We leave the integration of counting tasks into spatial benchmarks as future work.

Similarly, we exclude tasks that rely solely on distance measurements. Although distance is a fundamental spatial quantity, it is difficult to define consistently across different environments, object scales, and robot embodiments. Absolute distances can vary significantly between indoor and outdoor scenes, small and large objects, or different robot perspectives, making them hard to normalize or interpret in a general way. Moreover, distance alone often lacks the relational semantics required for higher-level reasoning—for example, understanding that an object is behind, above, or in front of others. ROBOSPATIAL instead focuses on spatial relationships that are more invariant, interpretable, and transferable across diverse robotic scenarios.

That said, the data generation pipeline is general and could readily support auxiliary tasks involving object counting or distance estimation if desired. These metrics may serve as useful complements in future extensions of the benchmark or as auxiliary supervision signals in model training.

B.3. Object Grounding Dataset

To support accurate spatial understanding, we generate an auxiliary dataset for object grounding. Many spatial reasoning tasks assume that the model can correctly identify which object is being referred to in the scene. However, in practice, this can be a major source of error—especially in cluttered environments or when multiple instances of the similar object type are present.

The grounding dataset provides direct supervision to help models learn to associate text descriptions with specific objects in the image. For each image, we include a set of object descriptions (e.g., “the keyboard” or “the chair”) paired with the corresponding 2D bounding box of the object in the image. These 2D boxes are projected from the annotated 3D bounding boxes using camera intrinsics and extrinsics.

A total of 100k grounding QA pairs are generated and used during training to reduce reference ambiguity and improve object identification accuracy in spatial tasks. While not part of the main spatial reasoning taxonomy, grounding accuracy is a prerequisite for answering spatial questions correctly, and we find that including this data helps reduce errors caused by incorrect object identification.

B.4. Dataset Generation Details

The dataset generation pipeline is detailed in the main text (subsection 3.2), which introduces a two-stage process for computing 3D spatial relationships and projecting them into 2D image space. Here, we expand on implementation details not covered in the main paper and provide clarification on the reasoning logic used in spatial annotation.

Category	Dataset	Split	Scans	Images	Configuration Q	Context Q	Compatibility Q
Indoor	Matterport3D [4]	Train	1859 scans	236243	298439	298439	298439
		Validation	10 scans	200	200	200	200
	ScanNet [8]	Train	1514 scans	280402	299039	299039	299039
		Validation	12 scans	400	400	400	400
	3RScan [56]	Train	1543 scans	366755	298839	298839	298839
		Validation	18 scans	400	400	400	400
Tabletop	HOPE [54]	Train	60 scenes	50050	36817	36817	36817
		Validation	47 scenes	235	500	500	500
	GraspNet-1B [12]	Train	130 scenes	25620	36817	36817	36817
		Validation	30 scenes	120	500	500	500

Table 7. Full dataset statistics for indoor and tabletop datasets.

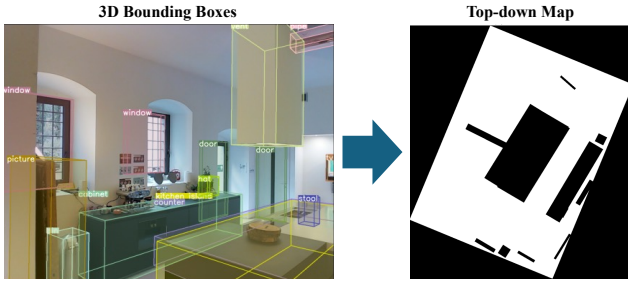


Figure 5. An example of generated top-down map of the image from 3D bounding boxes.

Reference Frame Annotation. For each spatial configuration question, we label relationships from three perspectives: ego-centric (camera view), object-centric (based on object heading), and world-centric (aligned with the dataset’s global frame). To compute object-centric directions, we use the heading vector of each oriented 3D bounding box to define the “front” of the object. Left, right, behind, and front relations are then assigned accordingly. World-centric annotations modify vertical relationships (above/below) using global z -coordinates to reflect elevation.

Surface Detection and Free Space Sampling. To identify support surfaces such as tables, counters, or floors, we use GPT-4o to select candidate objects that are likely to support placement. A top-down occupancy map is constructed from bounding boxes in the scene Fig. 5. We sample 3D points in unoccupied regions and project them into the image plane for spatial context tasks. Points are filtered via occlusion checks using raycasting, ensuring sampled points are visible and unobstructed.

Compatibility Check and Object Placement. For spatial compatibility, we simulate placing a virtual object bounding box at candidate locations. The placement must fit without intersecting other objects and must allow a clearance of at least 10 cm in all axes. We allow in-plane rotation and translation to test flexible placement. This provides a binary label (True/False) indicating whether the object can be compatibly placed in the region.

Output Format. Though ROBOSPATIAL uses point prediction

for ease of integration with robot setups, the pipeline also supports mask-based outputs and can be extended in future work.

C. Implementation Details

C.1. Model Training

We further explain the training details for all 2D and 3D VLMs trained on ROBOSPATIAL. For all models, we perform instruction tuning using the model weights from public repositories. All training is done using 8 Nvidia H100 GPUs, with the training time between 20 and 40 hours.

C.2. Model Setup

VILA [30] We initialize the model from Efficient-Large-Model/Llama-3-VILA1.5-8B on Hugging Face. We use the fine-tuning script from the VILA GitHub repository to train the model using the default hyperparameters.

LLaVA-NeXT [35] We initialize the model from lmms-lab/llama3-llava-next-8b on Hugging Face. We use the LLaVA-Next fine-tuning script from the LLaVA-Next repository using the default hyperparameters.

SpaceLLaVA [5] As official code and weights for SpatialVLM [5] is not released, we use a community implementation which is endorsed by SpatialVLM [5] authors. We initialize the model from remyxai/SpaceLLaVA from Hugging Face. We use LLaVA-1.5 finetuning script from LLaVa [34] repository using the default hyperparameters.

RoboPoint [62] We initialize the model from wentao-yuan/robopoint-v1-vicuna-v1.5-13b on Hugging Face. We use the fine-tuning script provided in the RoboPoint [62] GitHub repository to train the model using the default hyperparameters.

3D-LLM [18] We initialize the model using the pre-train_blip2_sam_flant5x1_v2.pth checkpoint downloaded from the official GitHub repository. Since the model requires preprocessing of multiview images, we follow the author’s pipeline to process multiview images from the environments. Because the model does not accept image input, we append the following text in front of the question to ensure the model understands the perspective from which the question is being asked: “I am facing ANCHOR OBJECT.” We use the default hyperparameters and train the model

Model	Indoor			Tabletop			Average		
	Ego-centric	Object-centric	World-centric	Ego-centric	Object-centric	World-centric	Indoor	Tabletop	Total
<i>Open-source VLMs</i>									
2D VLMs									
VILA [30]	55.9	40.5	32.9	43.6	39.7	28.9	43.1	37.4	40.2
+ROBOSPATIAL	74.3 \uparrow	57.8 \uparrow	62.3 \uparrow	70.3 \uparrow	58.1 \uparrow	60.3 \uparrow	64.8 \uparrow	62.9 \uparrow	63.9 \uparrow
LLaVA-Next [35]	35.2	24.3	34.7	36.4	28.5	22.7	31.4	29.2	30.3
+ROBOSPATIAL	75.4 \uparrow	54.1 \uparrow	68.8 \uparrow	67.9 \uparrow	54.7 \uparrow	58.9 \uparrow	60.4 \uparrow	60.5 \uparrow	60.5 \uparrow
SpaceLLaVA [5]	40.6	36.0	30.1	52.3	32.8	53.5	38.9	46.2	43.6
+ROBOSPATIAL	78.5\uparrow	60.6\uparrow	64.3 \uparrow	73.0 \uparrow	49.5 \uparrow	68.3 \uparrow	67.8 \uparrow	63.6 \uparrow	65.7 \uparrow
RoboPoint [62]	41.9	36.2	40.7	46.2	30.5	37.9	39.6	38.2	38.9
+ROBOSPATIAL	76.4 \uparrow	58.3 \uparrow	78.3 \uparrow	76.7\uparrow	62.6 \uparrow	71.0 \uparrow	71.0 \uparrow	70.1 \uparrow	70.6 \uparrow
3D VLMs									
3D-LLM [18]	28.9	38.3	45.6	38.9	35.7	52.6	37.6	42.4	40.0
+ROBOSPATIAL	60.7 \uparrow	52.1 \uparrow	76.5 \uparrow	57.9 \uparrow	62.8\uparrow	77.3 \uparrow	63.1 \uparrow	66.0 \uparrow	64.6 \uparrow
LEO [20]	46.9	30.6	48.2	41.4	34.3	55.4	41.9	43.7	42.8
+ROBOSPATIAL	68.1 \uparrow	71.6 \uparrow	79.6\uparrow	71.4 \uparrow	60.2 \uparrow	80.5\uparrow	73.1\uparrow	70.7\uparrow	71.9\uparrow
<i>Not available for fine-tuning</i>									
2D VLMs									
Molmo [9]	50.4	50.8	47.6	64.4	33.6	53.8	49.6	50.6	50.1
GPT-4o [42]	52.9	38.7	56.3	62.5	30.7	63.7	49.3	52.3	50.8

Table 8. Results of per frame accuracy of existing 2D/3D VLMs on a ROBOSPATIAL-Val. All methods, for all tasks, perform better (\uparrow) when fine-tuned on ROBOSPATIAL. The best result for each column is bolded.

Annotation Size	100K	300K	900k (Default)	1.8M	3M (Full)
LLaVA-Next [35]	38.1	46.7	60.5	65.8	72.4

Table 9. Results of scaling experiment on LLaVA-Next [35] with varied number of spatial relationship annotations. Average accuracy on ROBOSPATIAL-Val is reported.

	MMMU _{val}	MME _p	MME _c	MMBench _{dev}
LLaVA-NeXT	39.4	1561.8	305.4	71.6
+ROBOSPATIAL	39.8	1604.5	293.2	71.6

Table 10. Evaluation on general-purpose multimodal benchmarks (MMMU, MME, MMBench) to assess whether training on ROBOSPATIAL affects commonsense and factual reasoning.

	Base	Auxiliary	ROBOSPATIAL	Both
LLaVA-NeXT	30.3	32.4	51.8	60.5

Table 11. Ablation study evaluating the impact of the auxiliary grounding dataset on ROBOSPATIAL-Val.

for 20 epochs per the author’s guidelines. We choose the best model based on validation accuracy.

LEO [20] We initialize the model from the sft_noact.pth checkpoint downloaded from the official GitHub repository.

Since LEO supports dual image and 3D point cloud input, we input both of them and modify the question as in 3D-LLM. We use the default hyperparameters and train the model for 10 epochs per the author’s guidelines, and choose the best model based on

validation accuracy.

We could not fine-tune Molmo [9] from allenai/Molmo-7B-D-0924 or GPT-4o [42] from the gpt-4o-2024-08-06 API due to the unavailability of the fine-tuning script at the time of this work, thus we use them as a zero-shot baselines.

D. More Results

D.1. Accuracy Per Reference Frame

We show the results per frame in Tab. 8 for ROBOSPATIAL-Val. From the results, we can see a distinct difference between 2D and 3D VLMs in understanding the world-centric frame before training with ROBOSPATIAL. Baseline 2D VLMs have trouble understanding the world-centric frame, which involves understanding elevation, while 3D VLMs comparatively excel at it. Furthermore, we can see that since baseline 3D VLMs are trained on point clouds without information of perspective, their accuracy in ego-centric and object-centric frames is lower. However, with ROBOSPATIAL training, we were able to teach the 3D VLMs to think in a certain frame, thus considerably improving their performance on ego-centric and object-centric frames. However, we hypothesize that, due to their design—specifically, the lack of a means to visually inject perspective information since they require complete 3D point clouds—3D VLMs still lag behind 2D VLMs on ego-centric and object-centric frames.

D.2. Data Scaling

In Tab. 9, we experiment with scaling the number of annotations while keeping images fixed. We found that even though the number of images stays consistent, increasing the number of annotations can improve performance. For future work, we plan to apply

the data generation pipeline to a diverse set of indoor and tabletop environments to further improve the performance of the models.

D.3. Commonsense Knowledge Retention

To ensure that training on ROBOSPATIAL does not degrade a model’s general reasoning or commonsense capabilities, we evaluate the RoboSpatial-trained model on a suite of standard multimodal benchmarks: MMMU [63], MME [14], and MM-Bench [36]. As shown in Table 10, the ROBOSPATIAL-trained model maintains or slightly improves performance across all benchmarks, suggesting that spatial fine-tuning preserves broader knowledge capabilities.

D.4. Ablation of the Auxiliary Grounding Dataset

As shown in Table 11, training on the auxiliary dataset alone yields a small improvement over the base model (+2.1), but it falls far short of the gains achieved with ROBOSPATIAL, which is explicitly designed to teach spatial reasoning. This confirms that grounding supervision alone is insufficient for spatial understanding. However, combining both datasets leads to the best performance, suggesting that improving object localization can complement spatial supervision when jointly trained.

D.5. Robot Experiments Details

D.5.1. Robot Setup

For picking, we find which object the point maps to using SAM 2 [45] and execute the picking behavior on that object. For placing, we simply compute the 3D coordinate based on the depth value at that pixel and place the object at that coordinate. There were no failures due to cuRobo [52] failing. The experiments were purposely designed to consist of behaviors that our robot system can handle in order to avoid introducing irrelevant factors. The picking behavior consists of computing a top-down grasp pose and reaching it with cuRobo [52]. To compute the grasp pose:

1. We estimate the major axis of the object’s point cloud in top-down view using PCA.
2. The grasp orientation is orthogonal to the major axis.
3. The grasp height is based on the highest point in the object’s point cloud minus an offset of 3cm. This heuristic ensures the system can grip long objects.

The placing behavior is the same as picking, except that an area within 5cm of the placement coordinate is used as the point cloud for estimating orientation and height, and a vertical height offset is added to account for the height at which the object was picked.

D.5.2. Additional Results

We present additional results from the robot experiments in Fig. 6. We observe that models trained with ROBOSPATIAL consistently outperform baseline models in most cases, even though the prompt is not optimized for ROBOSPATIAL-trained models. This demonstrates that the power of VLMs enables templated language to generalize to language unseen during training while maintaining spatial understanding capabilities. However, even with ROBOSPATIAL training, the models struggle with understanding stacked items, indicating a need for further data augmentation with diverse layouts. In a few cases, ROBOSPATIAL training adversely affects performance, especially with RoboPoint [62]. We hypothesize that mixing the dataset with RoboPoint training data and RO-

BOSPATIAL training data may lead to unforeseen side effects, particularly in grounding objects. Nevertheless, we demonstrate that ROBOSPATIAL training enhances VLM’s spatial understanding in real-life robotics experiments, even with freeform language.

D.6. More Qualitative Examples

Fig. 7 present additional qualitative comparisons between models trained on ROBOSPATIAL. The findings demonstrate that models trained on ROBOSPATIAL consistently exhibit spatial understanding in the challenging ROBOSPATIAL-Home dataset, even outperforming closed models like GPT-4o [42]. However, we observed that object grounding is a crucial prerequisite for spatial understanding; the improvement is often hindered by the model’s inability to ground objects in cluttered scenes, where GPT-4o performs more effectively. Additionally, we show that the ROBOSPATIAL-trained model successfully generalizes to unseen spatial relationships in BLINK-Spatial [15], including those involving distance, such as “touching.”







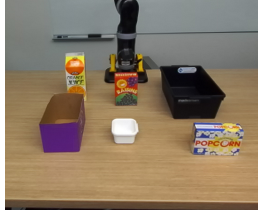
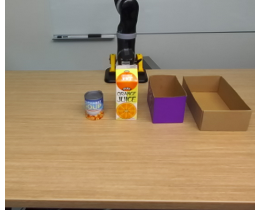






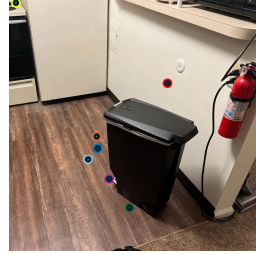
	<p>Question: pick lone object</p> <table><tr><td>LLaVa-Next [35]</td><td>×</td></tr><tr><td>LLaVa-Next-FT [35]</td><td>✓</td></tr><tr><td>RoboPoint [62]</td><td>×</td></tr><tr><td>RoboPoint-FT [62]</td><td>✓</td></tr><tr><td>Molmo [9]</td><td>✓</td></tr><tr><td>GPT-4o [42]</td><td>×</td></tr></table>	LLaVa-Next [35]	×	LLaVa-Next-FT [35]	✓	RoboPoint [62]	×	RoboPoint-FT [62]	✓	Molmo [9]	✓	GPT-4o [42]	×		<p>Question: Is there room to slot the pancake mix in the middle of the row of boxes</p> <table><tr><td>LLaVa-Next [35]</td><td>✓</td></tr><tr><td>LLaVa-Next-FT [35]</td><td>✓</td></tr><tr><td>RoboPoint [62]</td><td>×</td></tr><tr><td>RoboPoint-FT [62]</td><td>✓</td></tr><tr><td>Molmo [9]</td><td>✓</td></tr><tr><td>GPT-4o [42]</td><td>✓</td></tr></table>	LLaVa-Next [35]	✓	LLaVa-Next-FT [35]	✓	RoboPoint [62]	×	RoboPoint-FT [62]	✓	Molmo [9]	✓	GPT-4o [42]	✓
LLaVa-Next [35]	×																										
LLaVa-Next-FT [35]	✓																										
RoboPoint [62]	×																										
RoboPoint-FT [62]	✓																										
Molmo [9]	✓																										
GPT-4o [42]	×																										
LLaVa-Next [35]	✓																										
LLaVa-Next-FT [35]	✓																										
RoboPoint [62]	×																										
RoboPoint-FT [62]	✓																										
Molmo [9]	✓																										
GPT-4o [42]	✓																										
	<p>Question: Is there space in the white container for the orange juice box</p> <table><tr><td>LLaVa-Next [35]</td><td>×</td></tr><tr><td>LLaVa-Next-FT [35]</td><td>✓</td></tr><tr><td>RoboPoint [62]</td><td>×</td></tr><tr><td>RoboPoint-FT [62]</td><td>×</td></tr><tr><td>Molmo [9]</td><td>×</td></tr><tr><td>GPT-4o [42]</td><td>✓</td></tr></table>	LLaVa-Next [35]	×	LLaVa-Next-FT [35]	✓	RoboPoint [62]	×	RoboPoint-FT [62]	×	Molmo [9]	×	GPT-4o [42]	✓		<p>Question: alphabet soup fit in the purple box</p> <table><tr><td>LLaVa-Next [35]</td><td>✓</td></tr><tr><td>LLaVa-Next-FT [35]</td><td>×</td></tr><tr><td>RoboPoint [62]</td><td>✓</td></tr><tr><td>RoboPoint-FT [62]</td><td>✓</td></tr><tr><td>Molmo [9]</td><td>×</td></tr><tr><td>GPT-4o [42]</td><td>✓</td></tr></table>	LLaVa-Next [35]	✓	LLaVa-Next-FT [35]	×	RoboPoint [62]	✓	RoboPoint-FT [62]	✓	Molmo [9]	×	GPT-4o [42]	✓
LLaVa-Next [35]	×																										
LLaVa-Next-FT [35]	✓																										
RoboPoint [62]	×																										
RoboPoint-FT [62]	×																										
Molmo [9]	×																										
GPT-4o [42]	✓																										
LLaVa-Next [35]	✓																										
LLaVa-Next-FT [35]	×																										
RoboPoint [62]	✓																										
RoboPoint-FT [62]	✓																										
Molmo [9]	×																										
GPT-4o [42]	✓																										
	<p>Question: pick object behind the middle container</p> <table><tr><td>LLaVa-Next [35]</td><td>×</td></tr><tr><td>LLaVa-Next-FT [35]</td><td>✓</td></tr><tr><td>RoboPoint [62]</td><td>✓</td></tr><tr><td>RoboPoint-FT [62]</td><td>×</td></tr><tr><td>Molmo [9]</td><td>×</td></tr><tr><td>GPT-4o [42]</td><td>×</td></tr></table>	LLaVa-Next [35]	×	LLaVa-Next-FT [35]	✓	RoboPoint [62]	✓	RoboPoint-FT [62]	×	Molmo [9]	×	GPT-4o [42]	×		<p>Question: pick shortest object</p> <table><tr><td>LLaVa-Next [35]</td><td>×</td></tr><tr><td>LLaVa-Next-FT [35]</td><td>✓</td></tr><tr><td>RoboPoint [62]</td><td>✓</td></tr><tr><td>RoboPoint-FT [62]</td><td>✓</td></tr><tr><td>Molmo [9]</td><td>✓</td></tr><tr><td>GPT-4o [42]</td><td>✓</td></tr></table>	LLaVa-Next [35]	×	LLaVa-Next-FT [35]	✓	RoboPoint [62]	✓	RoboPoint-FT [62]	✓	Molmo [9]	✓	GPT-4o [42]	✓
LLaVa-Next [35]	×																										
LLaVa-Next-FT [35]	✓																										
RoboPoint [62]	✓																										
RoboPoint-FT [62]	×																										
Molmo [9]	×																										
GPT-4o [42]	×																										
LLaVa-Next [35]	×																										
LLaVa-Next-FT [35]	✓																										
RoboPoint [62]	✓																										
RoboPoint-FT [62]	✓																										
Molmo [9]	✓																										
GPT-4o [42]	✓																										
	<p>Question: place object in container behind popcorn</p> <table><tr><td>LLaVa-Next [35]</td><td>×</td></tr><tr><td>LLaVa-Next-FT [35]</td><td>✓</td></tr><tr><td>RoboPoint [62]</td><td>✓</td></tr><tr><td>RoboPoint-FT [62]</td><td>✓</td></tr><tr><td>Molmo [9]</td><td>×</td></tr><tr><td>GPT-4o [42]</td><td>×</td></tr></table>	LLaVa-Next [35]	×	LLaVa-Next-FT [35]	✓	RoboPoint [62]	✓	RoboPoint-FT [62]	✓	Molmo [9]	×	GPT-4o [42]	×		<p>Question: place the object inside the smallest box</p> <table><tr><td>LLaVa-Next [35]</td><td>×</td></tr><tr><td>LLaVa-Next-FT [35]</td><td>✓</td></tr><tr><td>RoboPoint [62]</td><td>✓</td></tr><tr><td>RoboPoint-FT [62]</td><td>✓</td></tr><tr><td>Molmo [9]</td><td>✓</td></tr><tr><td>GPT-4o [42]</td><td>×</td></tr></table>	LLaVa-Next [35]	×	LLaVa-Next-FT [35]	✓	RoboPoint [62]	✓	RoboPoint-FT [62]	✓	Molmo [9]	✓	GPT-4o [42]	×
LLaVa-Next [35]	×																										
LLaVa-Next-FT [35]	✓																										
RoboPoint [62]	✓																										
RoboPoint-FT [62]	✓																										
Molmo [9]	×																										
GPT-4o [42]	×																										
LLaVa-Next [35]	×																										
LLaVa-Next-FT [35]	✓																										
RoboPoint [62]	✓																										
RoboPoint-FT [62]	✓																										
Molmo [9]	✓																										
GPT-4o [42]	×																										
	<p>Question: can the robot directly pick the red orange peaches can without disturbing other objects?</p> <table><tr><td>LLaVa-Next [35]</td><td>✓</td></tr><tr><td>LLaVa-Next-FT [35]</td><td>✓</td></tr><tr><td>RoboPoint [62]</td><td>×</td></tr><tr><td>RoboPoint-FT [62]</td><td>×</td></tr><tr><td>Molmo [9]</td><td>✓</td></tr><tr><td>GPT-4o [42]</td><td>✓</td></tr></table>	LLaVa-Next [35]	✓	LLaVa-Next-FT [35]	✓	RoboPoint [62]	×	RoboPoint-FT [62]	×	Molmo [9]	✓	GPT-4o [42]	✓		<p>Question: is there an object that is not in a stack?</p> <table><tr><td>LLaVa-Next [35]</td><td>✓</td></tr><tr><td>LLaVa-Next-FT [35]</td><td>✓</td></tr><tr><td>RoboPoint [62]</td><td>✓</td></tr><tr><td>RoboPoint-FT [62]</td><td>✓</td></tr><tr><td>Molmo [9]</td><td>✓</td></tr><tr><td>GPT-4o [42]</td><td>✓</td></tr></table>	LLaVa-Next [35]	✓	LLaVa-Next-FT [35]	✓	RoboPoint [62]	✓	RoboPoint-FT [62]	✓	Molmo [9]	✓	GPT-4o [42]	✓
LLaVa-Next [35]	✓																										
LLaVa-Next-FT [35]	✓																										
RoboPoint [62]	×																										
RoboPoint-FT [62]	×																										
Molmo [9]	✓																										
GPT-4o [42]	✓																										
LLaVa-Next [35]	✓																										
LLaVa-Next-FT [35]	✓																										
RoboPoint [62]	✓																										
RoboPoint-FT [62]	✓																										
Molmo [9]	✓																										
GPT-4o [42]	✓																										
	<p>Question: can the macaroni and cheese be placed on top of cheez-it without touching other objects?</p> <table><tr><td>LLaVa-Next [35]</td><td>×</td></tr><tr><td>LLaVa-Next-FT [35]</td><td>×</td></tr><tr><td>RoboPoint [62]</td><td>✓</td></tr><tr><td>RoboPoint-FT [62]</td><td>✓</td></tr><tr><td>Molmo [9]</td><td>×</td></tr><tr><td>GPT-4o [42]</td><td>✓</td></tr></table>	LLaVa-Next [35]	×	LLaVa-Next-FT [35]	×	RoboPoint [62]	✓	RoboPoint-FT [62]	✓	Molmo [9]	×	GPT-4o [42]	✓		<p>Question: is there space to place one of the cans on the cheez-it box?</p> <table><tr><td>LLaVa-Next [35]</td><td>×</td></tr><tr><td>LLaVa-Next-FT [35]</td><td>×</td></tr><tr><td>RoboPoint [62]</td><td>×</td></tr><tr><td>RoboPoint-FT [62]</td><td>×</td></tr><tr><td>Molmo [9]</td><td>×</td></tr><tr><td>GPT-4o [42]</td><td>×</td></tr></table>	LLaVa-Next [35]	×	LLaVa-Next-FT [35]	×	RoboPoint [62]	×	RoboPoint-FT [62]	×	Molmo [9]	×	GPT-4o [42]	×
LLaVa-Next [35]	×																										
LLaVa-Next-FT [35]	×																										
RoboPoint [62]	✓																										
RoboPoint-FT [62]	✓																										
Molmo [9]	×																										
GPT-4o [42]	✓																										
LLaVa-Next [35]	×																										
LLaVa-Next-FT [35]	×																										
RoboPoint [62]	×																										
RoboPoint-FT [62]	×																										
Molmo [9]	×																										
GPT-4o [42]	×																										
	<p>Question: place on the object to the left of macaroni and cheese</p> <table><tr><td>LLaVa-Next [35]</td><td>×</td></tr><tr><td>LLaVa-Next-FT [35]</td><td>✓</td></tr><tr><td>RoboPoint [62]</td><td>✓</td></tr><tr><td>RoboPoint-FT [62]</td><td>✓</td></tr><tr><td>Molmo [9]</td><td>✓</td></tr><tr><td>GPT-4o [42]</td><td>×</td></tr></table>	LLaVa-Next [35]	×	LLaVa-Next-FT [35]	✓	RoboPoint [62]	✓	RoboPoint-FT [62]	✓	Molmo [9]	✓	GPT-4o [42]	×		<p>Question: pick the highest object on the stack of two objects</p> <table><tr><td>LLaVa-Next [35]</td><td>×</td></tr><tr><td>LLaVa-Next-FT [35]</td><td>×</td></tr><tr><td>RoboPoint [62]</td><td>×</td></tr><tr><td>RoboPoint-FT [62]</td><td>×</td></tr><tr><td>Molmo [9]</td><td>×</td></tr><tr><td>GPT-4o [42]</td><td>×</td></tr></table>	LLaVa-Next [35]	×	LLaVa-Next-FT [35]	×	RoboPoint [62]	×	RoboPoint-FT [62]	×	Molmo [9]	×	GPT-4o [42]	×
LLaVa-Next [35]	×																										
LLaVa-Next-FT [35]	✓																										
RoboPoint [62]	✓																										
RoboPoint-FT [62]	✓																										
Molmo [9]	✓																										
GPT-4o [42]	×																										
LLaVa-Next [35]	×																										
LLaVa-Next-FT [35]	×																										
RoboPoint [62]	×																										
RoboPoint-FT [62]	×																										
Molmo [9]	×																										
GPT-4o [42]	×																										

Figure 6. Additional robot experiments. A green check mark indicates that the model answered correctly. The -FT suffix denotes a model trained with ROBOSPATIAL. The questions are purposely not cleaned to reflect realistic language inputs.



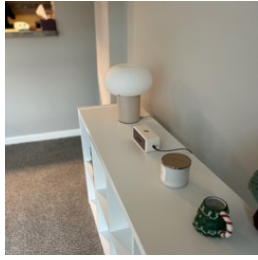
Question: Pinpoint several points within the vacant space situated to the left of the pot.

Answer	
LLaVa-Next [35]	●
LLaVa-Next-FT [35]	●
RoboPoint [62]	●
RoboPoint-FT [62]	●
Molmo [9]	●
GPT-4o [42]	●



Question: Pinpoint several points within the vacant space situated behind the trash bin.

Answer	
LLaVa-Next [35]	●
LLaVa-Next-FT [35]	●
RoboPoint [62]	●
RoboPoint-FT [62]	●
Molmo [9]	●
GPT-4o [42]	●



Question: Can the lamp fit in front of the shelf?

Answer	Yes
LLaVa-Next [35]	×
LLaVa-Next-FT [35]	✓
RoboPoint [62]	×
RoboPoint-FT [62]	✓
Molmo [9]	×
GPT-4o [42]	×



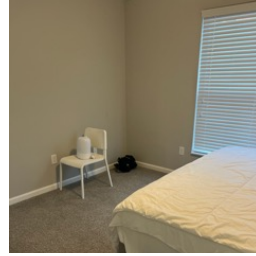
Question: Can the pot fit above the fridge?

Answer	Yes
LLaVa-Next [35]	×
LLaVa-Next-FT [35]	✓
RoboPoint [62]	×
RoboPoint-FT [62]	✓
Molmo [9]	×
GPT-4o [42]	×



Question: Is the lamp above the shelf?

Answer	Yes
LLaVa-Next [35]	×
LLaVa-Next-FT [35]	✓
RoboPoint [62]	×
RoboPoint-FT [62]	✓
Molmo [9]	×
GPT-4o [42]	✓



Question: Is the chair behind the bed?

Answer	Yes
LLaVa-Next [35]	×
LLaVa-Next-FT [35]	✓
RoboPoint [62]	×
RoboPoint-FT [62]	✓
Molmo [9]	×
GPT-4o [42]	×



Question: Is the dining table touching the donut?

Answer	Yes
LLaVa-Next [35]	×
LLaVa-Next-FT [35]	✓
RoboPoint [62]	×
RoboPoint-FT [62]	✓
Molmo [9]	×
GPT-4o [42]	×



Question: Is the couch under the suitcase?

Answer	Yes
LLaVa-Next [35]	×
LLaVa-Next-FT [35]	✓
RoboPoint [62]	×
RoboPoint-FT [62]	✓
Molmo [9]	×
GPT-4o [42]	×

Figure 7. Qualitative results on spatial reasoning benchmarks. The -FT suffix denotes a model trained with ROBOSPATIAL. The first three rows show examples from ROBOSPATIAL-Home, covering spatial context, spatial compatibility, and spatial configuration. For spatial context questions, only the first predicted point from each model is shown. The fourth row shows generalization to unseen spatial relationships on the Blink-Spatial [15] dataset, demonstrating that the ROBOSPATIAL-trained model can transfer to unseen relationships.