ROBOSPATIAL: TEACHING SPATIAL UNDERSTANDING TO 2D AND 3D VISION-LANGUAGE MODELS FOR ROBOTICS

Chan Hee Song[1*†] **Valts Blukis**[2] **Jonathan Tremblay**[2] **Stephen Tyree**[2]
Yu Su[1] **Stan Birchfield**[2]
[1]The Ohio State University [2]NVIDIA
Website: https://chanh.ee/RoboSpatial

ABSTRACT

Spatial understanding is essential for robots to perceive, reason about, and interact with their environments. However, current visual language models often rely on general-purpose image datasets that lack robust spatial scene understanding and reference frame comprehension (ego-, world-, or object-centric). To address this gap, we introduce ROBOSPATIAL, a large-scale dataset of real indoor and tabletop environments captured via egocentric images and 3D scans. ROBOSPATIAL provides 1M images, 5k 3D scans, and 3M annotated spatial relationships, enabling both 2D and 3D spatial reasoning. Models trained on ROBOSPATIAL outperform baselines on tasks including spatial affordance prediction, spatial relationship prediction, and robot manipulation.

1 INTRODUCTION

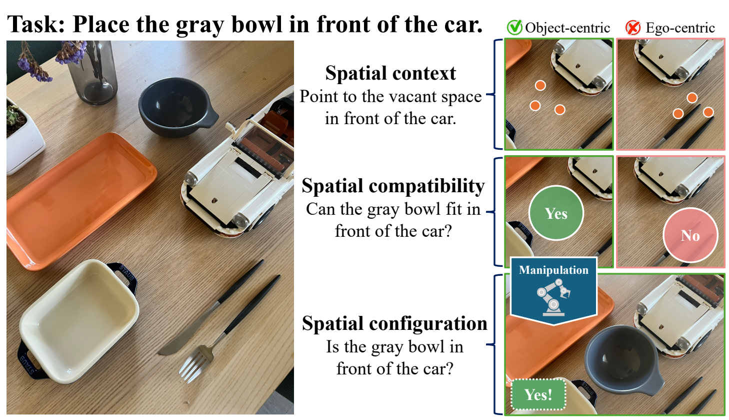


Figure 1: ROBOSPATIAL dataset facilitates 3D spatial reasoning for robot manipulation. This illustration demonstrates how a model trained on ROBOSPATIAL enables human-aligned spatial reasoning within the correct reference frame, supporting task grounding, planning, and detection for manipulation tasks.

Recent advances in vision-language models (VLMs) have begun to bridge the gap between computer vision and robotics control. VLMs trained directly on robot manipulation data now enable robots to process both visual inputs and task descriptions in real-world settings Collaboration et al. (2023). Similarly, generic VLMs have been used to describe robotics scenes for specific tasks Fang et al. (2024), while large language models (LLMs) have demonstrated utility in generating robot code

*Correspondence to: song.1855@osu.edu
†This is a short version of the paper accepted to CVPR 2025 (Oral).

Dataset	3D scans	Embodied	Ref. frames	Compatibility	Domain	#Scans	#Images	#Spatial QAs
EmbSpatial-Bench Du et al. (2024)	✓	✓	✗	✗	Indoor	277	2k	4k
Visual Spatial Liu et al. (2023a)	✗	✗	✓	✗	MSCOCO	0	10k	10k
SpatialRGPT-Bench Cheng et al. (2024)	✗	✗	✗	✓	Indoor, AV	0	1.4k	1.4k
BLINK-Spatial Fu et al. (2024)	✗	✗	✓	✗	Generic	0	286	286
What's up Kamath et al. (2023)	✗	✗	✗	✗	Generic	0	5k	10k
Spatial-MM Shiri et al. (2024)	✗	✗	✓	✗	Generic	0	2.3k	2.3k
RoboSpatial	✓	✓	✓	✓	Indoor, tabletop	5k	1M	3M

Table 1: Comparison with other spatial reasoning datasets that include object-centric spatial relationships.

and planning high-level tasks Ahn et al. (2022); Singh et al. (2023); Liang et al. (2023); Song et al. (2023).

Despite successes in object recognition and scene description, current VLMs still lack nuanced spatial understanding Yamada et al. (2024); Kamath et al. (2023). For example, while a model might accurately describe a "bowl on the table," it struggles to reason about the optimal placement of the bowl in terms of accessibility, stability, or its fit among other objects. A significant challenge lies in the fact that existing training datasets do not capture the varied reference frames—first-person, object-centric, or global—required for robust real-world interactions.

Some recent works have attempted to improve spatial reasoning, such as SpatialVLM Chen et al. (2024) and SpatialRGPT Cheng et al. (2024), which focus on conceptual spatial relationships, or RoboPoint Yuan et al. (2024), which predicts grounded 2D coordinates. However, these models often rely on web images or synthetic data and thus fail to generalize to robot-captured images, which lack identifiable scale cues and real-world constraints. Similarly, although Molmo Deitke et al. (2024) shows promise for object-centric image-space pointing, it struggles with practical constraints, such as determining if an object can physically fit in a designated space.

Motivated by these limitations, we introduce RoboSpatial and RoboSpatial-Home, a training dataset and benchmark specifically designed to enhance spatial reasoning for robotic applications. Leveraging annotated indoor scene and tabletop RGBD data, we transform these into targeted question-answer pairs that probe critical spatial skills, including object-object relationships, object-space interactions, and object compatibility. Each question is posed from three distinct reference frames—ego-centric (the observer's viewpoint), object-centric, and world-centric—to better capture the complexity of spatial instructions. RoboSpatial comprises approximately 1M images, 5k 3D scans, and 3M annotated spatial relationships, making it well-suited for both 2D and 3D tasks (see Figure 1).

We validate our dataset by training state-of-the-art 2D and 3D VLMs, which demonstrate significant improvements in spatial reasoning over existing models. The enhanced models outperform prior approaches on our validation split (RoboSpatial-Val) and on additional downstream tasks, including RoboSpatial-Home, BLINK-Spatial Fu et al. (2024), SpatialBench Cai et al. (2025) and real-world robot manipulation. Our experiments further compare 2D and 3D VLM performance, underscoring the benefits of incorporating 3D-based training for robust spatial understanding in robotics.

2 Approach

Our approach centers on three core spatial relationships—configuration, context, and compatibility—that together form a nuanced framework for robotic spatial reasoning. These relationships guide our automated data generation pipeline for constructing RoboSpatial.

We define **spatial configuration** as the ability to interpret the relative positioning of objects; for example, determining if one object is to the left of an anchor object. This binary relationship is essential for navigation, manipulation, and interaction. **Spatial context** involves identifying specific points (in image coordinates) relative to an anchor object, such as determining where in free space an object may be placed. Here, we generate a top-down map from annotated 3D bounding boxes and sample candidate points based on object size, with answers provided as lists of 2D coordinates. Finally, **spatial compatibility** extends the context task by assessing whether a referenced object can physically fit within a designated region relative to the anchor. This task simulates object placement using bounding box sizes and yields binary answers. To enhance the model's ability to interpret

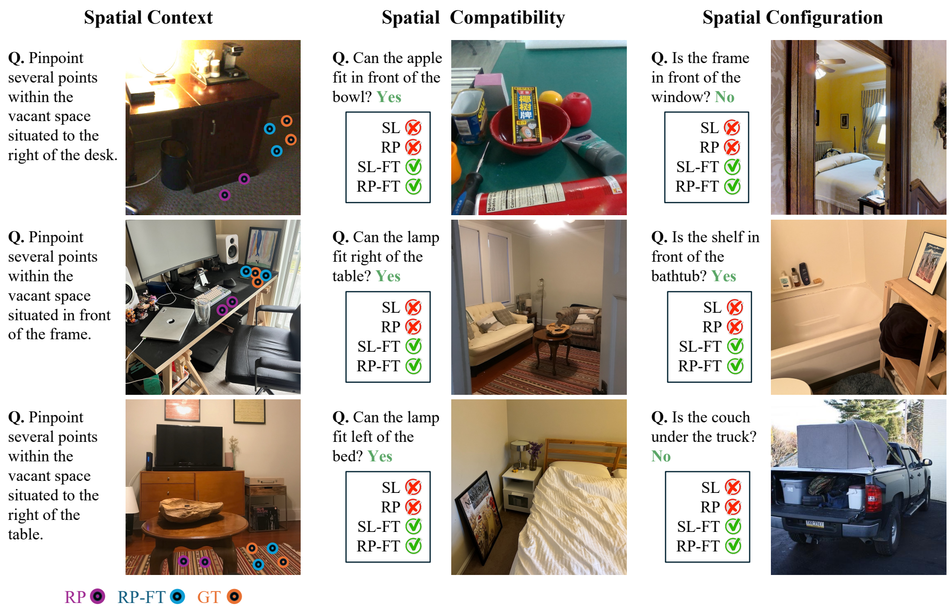


Figure 2: In-domain (RoboSpatial-Val, top) and out-of-domain (RoboSpatial-Home, BLINK Fu et al. (2024), middle and bottom) results for RoboSpatial-trained models. Two models shown: SL (SpaceLLaVA Chen et al. (2024)) and RP (RoboPoint Yuan et al. (2024)); the -FT suffix indicates fine-tuning on RoboSpatial. Correct answers in green. All images except bottom-right in the out-of-domain rows are from RoboSpatial-Home.

spatial instructions from different perspectives, each question-answer pair in RoboSpatial is posed from three distinct reference perspectives/frames: (a) **Ego-centric** from the observer's perspective at the camera pose, (b) **World-centric** grounded in a global world frame, and (c) **Object-centric** based on a reference frame attached to the focal object.

Our data generation pipeline minimizes human intervention through carefully constructed heuristics. Starting with a source dataset, $\mathcal{D}_s$, that provides RGB images, camera poses, text labels, and oriented 3D bounding boxes, we generate a new dataset $\mathcal{D}$ in which each datum $d_i = \langle I_i, q_i, a_i, l_i \rangle$ includes an image, a question, an answer, and a reference frame label (ego, world, or object). The process unfolds in two stages. First, in the *spatial relation extraction* stage, we automatically derive spatial relationships of the form $\langle I_i, a_i, t_i, s_i, r_i, l_i \rangle$, where I_i is the source image, a_i is the anchor object, t_i is the target object or a sampled point in free space, $r_i \in \{\textit{left}, \textit{right}, \textit{above}, \textit{below}, \textit{front}, \textit{behind}\}$ is the relation preposition, and $l_i \in \{\textit{ego}, \textit{world}, \textit{object}\}$ denotes the reference frame. For spatial configuration, an anchor object is paired with all other uniquely appearing objects in the image according to the specified direction and reference frame. For spatial context, we compute candidate placement points in free space via a top-down map, while for spatial compatibility we simulate placing an object using its bounding box size to determine feasibility.

In the second stage, *question-answer generation*, we transform these spatial relationships into template-based pairs following the structure "{object/space} {relationship} {anchor object} {reference frame}." This templating ensures that questions are unambiguous and that models rely on visual reasoning rather than linguistic commonsense. Additionally, we generate an auxiliary object-referring dataset using 2D bounding boxes to improve object grounding. Overall, our pipeline produces 3M spatial relationships—an order of magnitude more than previous datasets (see Table 1)—covering a comprehensive range of spatial reasoning tasks.

3 Implementation and Evaluation

We apply our data generation process to diverse datasets, including three scene datasets—ScanNet Dai et al. (2017), Matterport3D Chang et al. (2017), and 3RScan Wald

Table 2: Spatial reasoning and robot manipulation results. "R-" denotes ROBOSPATIAL. QA pairs are evaluated by average accuracy, and robot performance is reported as success rate.

Model	R-Test	R-Home	BLINK-Spatial	SpatialBench-Position	Robot
Open source – 2D					
LLaVA-NeXT	30.3	46.3	71.8	55.9	23.7
+ RoboSpatial	60.5	59.6	**79.0**	**70.6**	**52.6**
RoboPoint	38.9	53.4	63.6	44.1	44.7
+ RoboSpatial	70.6	**63.4**	70.6	64.7	46.2
Open source – 3D					
Embodied Generalist	42.8	29.8	N/A	N/A	N/A
+ RoboSpatial	**71.9**	43.8	N/A	N/A	N/A
Closed source					
Molmo	50.1	25.6	67.1	55.9	43.8
GPT-4o	50.8	47.0	76.2	**70.6**	46.9

et al. (2019)—and two tabletop datasets—HOPE Tyree et al. (2022) and GraspNet-1B Fang et al. (2020). Using 3D bounding boxes and embodied images from EmbodiedScan Wang et al. (2024b), we generate a large-scale spatial reasoning dataset with approximately 3M QA pairs, 5k 3D scans, and 1M images.

We evaluate a range of 2D and 3D vision-language models (VLMs). For 2D models, we compare base models (VILA-1.5-8B Lin et al. (2024) and LLaVA-NeXT-8B Liu et al. (2024)), specialized models (SpaceLLaVA-13B, RoboPoint-13B Yuan et al. (2024), and Molmo-7B Deitke et al. (2024)), and the closed-source GPT-4o OpenAI et al. (2024) (omitting models like SpatialRGPT Cheng et al. (2024) that rely on external object masks). For 3D models, we test 3D-LLM Hong et al. (2023) (which reconstructs 3D point clouds from multi-view images) and LEO Huang et al. (2024b) (which processes segmented 3D point clouds). We report both zero-shot and fine-tuned (on ROBOSPATIAL) performance (full results and details in the Appendix).

Spatial understanding is assessed across four benchmarks: ROBOSPATIAL-Val, ROBOSPATIAL-Test, BLINK Fu et al. (2024), and SpatialBench Cai et al. (2025), covering over 6,000 questions across binary (yes/no) and numeric (2D coordinate prediction) formats. For binary questions, we report accuracy; for numeric questions, we measure whether the model's prediction lies within the convex hull of reference points derived from scene geometry. These datasets span both in-domain and out-of-domain settings, capturing variation in visual environments and language formulations. To evaluate real-world grounding, we additionally deploy models on a Kinova Jaco robot tasked with spatially grounded pick-and-place manipulation. Here, predicted answers or coordinates are executed via a motion planning system, testing end-to-end spatial understanding in physical environments. Table 2 presents the main results.

4 RESULTS AND DISCUSSION

Our experiments demonstrate that training on ROBOSPATIAL substantially improves spatial reasoning across models. Models trained on ROBOSPATIAL more accurately align their predictions with intended reference frames, exhibiting a better understanding of spatial relations such as directionality and relative positioning. While existing models often struggle with ambiguous or underspecified spatial cues, ROBOSPATIAL-trained models infer appropriate placements by leveraging object geometry and contextual cues. Although the dataset is built on templated spatial relationships, the models generalize to novel prepositions by mapping principal 3D directions to corresponding linguistic terms. This training also enhances the ability to interpret nuanced, context-dependent reference frames-an essential capability for real-world spatial understanding. Additionally, while 3D VLMs tend to outperform 2D models due to access to depth information, 2D models remain highly sensitive to minor pixel-level inaccuracies, which can lead to significant misalignments when translated into 3D space for robot manipulation.

REFERENCES

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*, 2022.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, June 2024.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki

Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Mart'in-Mart'in, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open X-Embodiment: Robotic learning datasets and RT-X models. `https://arxiv.org/abs/2310.08864`, 2023.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Jen Dumas, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. EmbSpatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 346–355, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.33. URL `https://aclanthology.org/2024.acl-short.33`.

Jiafei Duan, Wilbert Pumacay, Nishanth Kumar, Yi Ru Wang, Shulin Tian, Wentao Yuan, Ranjay Krishna, Dieter Fox, Ajay Mandlekar, and Yijie Guo. AHA: A vision-language-model for detecting and reasoning over failures in robotic manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=JVkdSi7Ekg`.

Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11444–11453, 2020.

Kuan Fang, Fangchen Liu, Pieter Abbeel, and Sergey Levine. Moka: Open-world robotic manipulation through mark-based visual prompting. *Robotics: Science and Systems (RSS)*, 2024.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 148–166. Springer, 2024. doi: 10.1007/978-3-031-73337-6_9. URL `https://doi.org/10.1007/978-3-031-73337-6_9`.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. In *Advances in Neural Information Processing Systems*, 2023. NeurIPS.

Haoxu Huang, Fanqi Lin, Yingdong Hu, Shengjie Wang, and Yang Gao. Copa: General robotic manipulation through spatial constraints of parts with foundation models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9488–9495, 2024a. doi: 10.1109/IROS58592.2024.10801352.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024b.

Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. In *8th Annual Conference on Robot Learning*, 2024c. URL `https://openreview.net/forum?id=9iG3SEbMnL`.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6700–6709, 2019.

Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Proceedings of the 8th Conference on Robot Learning (CoRL)*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 06–09 Nov 2025. URL `https://proceedings.mlr.press/v270/kim25c.html`.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9493–9500, 2023. doi: 10.1109/ICRA48891.2023.10160591.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: On Pre-training for Visual Language Models . In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26679–26689, Los Alamitos, CA, USA, June 2024. IEEE Computer Society. doi: 10.1109/CVPR52733.2024.02520. URL `https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.02520`.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

Xiongkun Linghu, Jiangyong Huang, Xuesong Niu, Xiaojian Ma, Baoxiong Jia, and Siyuan Huang. Multi-modal situated reasoning in 3d scenes. In *Advances in Neural Information Processing Systems*, 2024. NeurIPS.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023a. doi: 10.1162/tacl_a_00566. URL https://aclanthology.org/2023.tacl-1.37.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, June 2024.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=IDJx97BC38.

Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Situational awareness matters in 3d vision language reasoning. In *CVPR*, 2024.

Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, Jacky Liang, Ishita Dasgupta, Annie Xie, Danny Driess, Ayzaan Wahid, Zhuo Xu, Quan Vuong, Tingnan Zhang, Tsang-Wei Edward Lee, Kuang-Huei Lee, Peng Xu, Sean Kirmani, Yuke Zhu, Andy Zeng, Karol Hausman, Nicolas Heess, Chelsea Finn, Sergey Levine, and Brian Ichter. Pivot: iterative visual prompting elicits actionable knowledge for vlms. In *Proceedings of the International Conference on Machine Learning (ICML)*, ICML'24. JMLR.org, 2024.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, Tobias Kreiman, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Navid Rajabi and Jana Kosecka. Towards grounded visual spatial reasoning in multi-modal vision language models, 2024. URL https://arxiv.org/abs/2308.09778.

Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S. Ryoo, and Tsung-Yu Lin. Learning to localize objects improves spatial reasoning in visual-llms. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12977–12987, 2024. doi: 10.1109/CVPR52733.2024.01233.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

Leonard Salewski, A. Sophia Koepke, Hendrik P. A. Lensch, and Zeynep Akata. Clevr-x: A visual reasoning dataset for natural language explanations. In *xxAI - Beyond explainable Artificial Intelligence*, pp. 85–104. Springer, 2022.

Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21440–21455, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.1195.

Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11523–11530, 2023. doi: 10.1109/ICRA48891.2023.10161317.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In Anna Korhonen, David Traum, and

Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6418–6428, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1644. URL https://aclanthology.org/P19-1644/.

Balakumar Sundaralingam, Siva Kumar Sastry Hari, Adam Fishman, Caelan Garrett, Karl Van Wyk, Valts Blukis, Alexander Millane, Helen Oleynikova, Ankur Handa, Fabio Ramos, et al. cuRobo: Parallelized collision-free minimum-jerk robot motion generation. *arXiv preprint arXiv:2310.17274*, 2023.

Emilia Szymanska, Mihai Dusmanu, Jan-Willem Buurlage, Mahdi Rad, and Marc Pollefeys. Space3D-Bench: Spatial 3D Question Answering Benchmark. In *European Conference on Computer Vision (ECCV) Workshops*, 2024.

Stephen Tyree, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Jeffrey Smith, and Stan Birchfield. 6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark. In *International Conference on Intelligent Robots and Systems (IROS)*, 2022.

Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. Gpt-4v(ision) for robotics: Multimodal task planning from human demonstration. *IEEE Robotics and Automation Letters*, 9(11):10567–10574, 2024. doi: 10.1109/LRA.2024.3477090.

Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Niessner. Rio: 3d object instance re-localization in changing indoor environments. In *Proceedings IEEE International Conference on Computer Vision (ICCV)*, 2019.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=cvaSru8LeO.

Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, Xihui Liu, Cewu Lu, Dahua Lin, and Jiangmiao Pang. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Youngsun Wi, Mark Van der Merwe, Pete Florence, Andy Zeng, and Nima Fazeli. Calamari: Contact-aware and language conditioned spatial action mapping for contact-rich manipulation. In *7th Annual Conference on Robot Learning*, 2023.

Liuchang Xu, Shuo Zhao, Qingming Lin, Luyao Chen, Qianqian Luo, Sensen Wu, Xinyue Ye, Hailin Feng, and Zhenhong Du. Evaluating large language models on spatial tasks: A multi-task benchmarking study, 2024. URL https://arxiv.org/abs/2408.14438.

Yutaro Yamada, Yihan Bao, Andrew K. Lampinen, Jungo Kasai, and Ilker Yildirim. Evaluating spatial understanding of large language models, 2024. URL https://arxiv.org/abs/2310.14540.

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction in robotics. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=GVX6jpZOhU.

Yue Zhang, Zhiyang Xu, Ying Shen, Parisa Kordjamshidi, and Lifu Huang. SPARTUN3d: Situated spatial understanding of 3d world in large language model. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FGMkSL8NR0.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao

Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 06–09 Nov 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

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

- Appendix A Related Works.
- Appendix B Limitations.
- Appendix C Dataset Details.
- Appendix D Implementation Details.
- Appendix E Full Results.

A Related Works

VLMs for Robotics. Vision-language models (VLMs) have emerged as pivotal tools in robotics, enabling systems to interpret and act upon complex visual and textual information. By integrating visual perception with language understanding, VLMs facilitate more intuitive human-robot interactions and enhance autonomous decision-making capabilities. Recent advancements have demonstrated the potential of VLMs in various robotic applications. For instance, vision-language-action models (VLAs) Kim et al. (2025); Zitkovich et al. (2023); Octo Model Team et al. (2024) enable robots to interpret and execute complex instructions and output executable robot actions. Additionally, VLMs like GPT-4v OpenAI et al. (2024) have been utilized for high-level task planning Wake et al. (2024), allowing robots to generate detailed action sequences from natural language instructions. Furthermore, VLMs have been used for keypoint/mask prediction Huang et al. (2024c); Wi et al. (2023); Nasiriany et al. (2024), error analysis Duan et al. (2025), grasp pose prediction Huang et al. (2024a). Despite these advancements, integrating VLMs Cai et al. (2025); Cheng et al. (2024); Yuan et al. (2024) into robotic systems presents challenges. One significant hurdle is the need for precise spatial reasoning to navigate and manipulate objects effectively. While VLMs excel in understanding and generating language, their ability to comprehend and reason about spatial relationships in dynamic environments remains limited Yamada et al. (2024); Xu et al. (2024); Wang et al. (2024a). Therefore, RoboSpatial aims to tackle this gap by presenting a large scale pretraining and evaluation setup for teaching spatial understanding to VLM for robotics.

Spatial Understanding with VLMs. Spatial understanding has been implicitly and explicitly part of various vision and question answering tasks Fu et al. (2024); Azuma et al. (2022); Jia et al. (2024); Suhr et al. (2019); Salewski et al. (2022); Krishna et al. (2017); Johnson et al. (2017); Hudson & Manning (2019). While many benchmarks and methods have been proposed, they often come with limitations: some focus exclusively on simulations Szymanska et al. (2024) or generic images Liu et al. (2023a); Rajabi & Kosecka (2024); Cheng et al. (2024); Chen et al. (2024); Fu et al. (2024); Kamath et al. (2023); Shiri et al. (2024); Ranasinghe et al. (2024), others are difficult to evaluate Szymanska et al. (2024); Du et al. (2024); Linghu et al. (2024), rely on complete 3D scans Zhang et al. (2025); Man et al. (2024); Ma et al. (2023); Linghu et al. (2024), or do not consider reference frames Zhang et al. (2025); Man et al. (2024); Ma et al. (2023); Linghu et al. (2024); Chen et al. (2024); Cheng et al. (2024); Fu et al. (2024); Ranasinghe et al. (2024). Furthermore, they often fail to address actionable, robotics-relevant spatial relationships such as spatial compatibility and context Du et al. (2024); Fu et al. (2024); Wang et al. (2024b); Shiri et al. (2024); Kamath et al. (2023); Linghu et al. (2024); Ranasinghe et al. (2024).

Inspired by prior works on spatial reasoning Liu et al. (2023a); Kamath et al. (2023)—where the impact of reference frames and spatial configurations was explored in generic images Lin et al. (2014); Hudson & Manning (2019)—we extend spatial understanding to a robotics-specific context with actionable spatial relationships such as spatial compatibility and spatial context. Our aim is to enable direct application to robotic workflows, such as task planning and verification.

To achieve this, we have developed and are planning to open-source a large-scale 2D/3D ready pretraining dataset, an automated data annotation pipeline, and trained models. We further show how our dataset can be used to teach spatial reasoning to a suite of vision-language models (VLMs) in in-domain and out-of-domain spatial reasoning datasets. We hope these resources lower the barrier to entry for exploring spatial understanding tailored to robotics.

B 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 Kim et al. (2025). 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 generalizes well to out-of-distribution benchmarks like BLINK, which include humans and animals—suggesting that the learned spatial representations are broadly transferable.

C DATASET DETAILS

C.1 DATASET STATISTICS

We provide the full dataset statistics in Table 3. 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 Table 7 and explain the results. Notably, HOPE Tyree et al. (2022) and GraspNet-1B Fang et al. (2020) 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.

C.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.

C.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.

C.4 Dataset Generation Details

The dataset generation pipeline is detailed in the main text (section 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.

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 Figure 3. 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.

Category	Dataset	Split	Scans	Images	Configuration Q	Context Q	Compatibility Q
Indoor	Matterport3D Chang et al. (2017)	Train	1859 scans	236243	298439	298439	298439
		Validation	10 scans	200	200	200	200
	ScanNet Dai et al. (2017)	Train	1514 scans	280402	299039	299039	299039
		Validation	12 scans	400	400	400	400
	3RScan Wald et al. (2019)	Train	1543 scans	366755	298839	298839	298839
		Validation	18 scans	400	400	400	400
Tabletop	HOPE Tyree et al. (2022)	Train	60 scenes	50050	36817	36817	36817
		Validation	47 scenes	235	500	500	500
	GraspNet-1B Fang et al. (2020)	Train	130 scenes	25620	36817	36817	36817
		Validation	30 scenes	120	500	500	500

Table 3: Full dataset statistics for indoor and tabletop datasets.

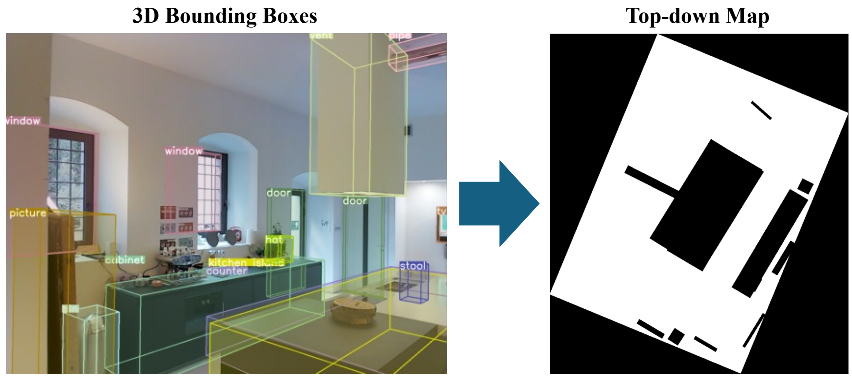


Figure 3: An example of generated top-down map of the image from 3D bounding boxes.

D Implementation Details

D.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.

VILA Lin et al. (2024) We initialize our 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 our model using the default hyperparameters.
LLaVA-NeXT Liu et al. (2024) We initialize our 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 Chen et al. (2024) As official code and weights for SpatialVLM Chen et al. (2024) is not released, we use a community implementation which is endorsed by SpatialVLM Chen et al. (2024) authors. We initialize our model from remyxai/SpaceLLaVA from Hugging Face. We use LLaVA-1.5 finetuning script from LLaVa Liu et al. (2023b) repository using the default hyperparameters.
RoboPoint Yuan et al. (2024) We initialize our model from wentao-yuan/robopoint-v1-vicuna-v1.5-13b on Hugging Face. We use the fine-tuning script provided in the RoboPoint Yuan et al. (2024) GitHub repository to train our model using the default hyperparameters.
3D-LLM Hong et al. (2023) We initialize our model using the pretrain_blip2_sam_flant5xl_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 our 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 for 20 epochs per the author's guidelines. We choose the best model based on validation accuracy.
LEO Huang et al. (2024b) We initialize our 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 Deitke et al. (2024) from allenai/Molmo-7B-D-0924 or GPT-4o OpenAI et al. (2024) 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.2 Robot Setup

For picking, we find which object the point maps to using SAM 2 Ravi et al. (2025) and execute our picking behavior on that object. For placing, we simply compute the 3D coordinate based on

Model	Indoor			Tabletop			Average		
	Configuration	Context	Compatibility	Configuration	Context	Compatibility	Indoor	Tabletop	Total
Open-source VLMs **2D VLMs**									
VILA Lin et al. (2024)	54.7	18.3	56.3	45.1	13.2	53.8	43.1	37.4	40.2
+RoboSpatial	71.4 ↑	45.9 ↑	77.2 ↑	71.8 ↑	43.7 ↑	73.3 ↑	64.8 ↑	62.9 ↑	63.9 ↑
LLaVA-NeXT Liu et al. (2024)	48.9	12.5	32.7	48.3	8.4	30.9	31.4	29.2	30.3
+RoboSpatial	69.3 ↑	41.3 ↑	70.5 ↑	70.7 ↑	44.8 ↑	66.1 ↑	60.4 ↑	60.5 ↑	60.5 ↑
SpaceLLaVA Chen et al. (2024)	52.6	15.3	49.0	66.5	12.2	60.1	38.9	46.2	43.6
+RoboSpatial	76.0 ↑	50.7 ↑	76.6 ↑	74.9 ↑	46.4 ↑	70.5 ↑	67.8 ↑	63.6 ↑	65.7 ↑
RoboPoint Yuan et al. (2024)	39.0	41.4	38.3	37.9	31.6	45.2	39.6	38.2	38.9
+RoboSpatial	72.2 ↑	**68.9** ↑	72.1 ↑	70.3 ↑	**61.7** ↑	78.4 ↑	71.0 ↑	70.1 ↑	70.6 ↑
3D VLMs									
3D-LLM Hong et al. (2023)	54.5	8.1	53.6	59.2	10.6	57.4	37.6	42.4	40.0
+RoboSpatial	76.3 ↑	35.4 ↑	77.5 ↑	76.2 ↑	46.8 ↑	75.0 ↑	63.1 ↑	66.0 ↑	64.6 ↑
LEO Huang et al. (2024b)	56.1	11.3	58.3	60.8	11.1	59.3	41.9	43.7	42.8
+RoboSpatial	**80.2** ↑	56.7 ↑	**82.5** ↑	**78.1** ↑	55.2 ↑	**78.9** ↑	**73.1** ↑	**70.7** ↑	**71.9** ↑
Not available for fine-tuning **2D VLMs**									
Molmo Deitke et al. (2024)	40.6	48.2	60.0	61.5	35.8	54.6	49.6	50.6	50.1
GPT-4o OpenAI et al. (2024)	63.5	25.1	59.4	62.3	27.9	66.8	49.3	52.3	50.8

Table 4: Results of existing 2D/3D VLMs on a held-out validation split (RoboSpatial-Val) of images and scans. All methods, for all tasks, perform better (↑) when fine-tuned on RoboSpatial. The best result for each column is bolded.

the depth value at that pixel and place the object at that coordinate. There were no failures due to cuRobo Sundaralingam et al. (2023) 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 Sundaralingam et al. (2023). 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.

E Full Results

E.1 Omitted Results in the Main Text

We show the full results in held-out test split in Table 4 and out-of-domain splits in Table 5.

E.2 Cross-Dataset Generalization

We evaluate the generalization capability of our method by testing it across different scene types—specifically, both indoor and tabletop scenes—to control for any bias in the annotations of the underlying datasets that make up our benchmark. We train on data derived from subsets of the datasets corresponding to one scene type (either indoor or tabletop) and test on held-out datasets from the other scene type, representing unseen environments. We expect that even when training on a subset of datasets, the performance on unseen scene types will improve if our method generalizes well. The results of this cross-dataset evaluation are shown in Table 6.

E.3 Data Scaling

In Table 7, 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 our data generation pipeline to a diverse set of indoor and tabletop environments to further improve the performance of our models.

Model	RoboSpatial-Home			BLINK	SpatialBench
	Configuration	Context	Compatibility	Accuracy	Accuracy
2D VLMs					
VILA Lin et al. (2024)	57.8	0.0	69.0	72.7	53.0
+RoboSpatial	65.9 ↑	15.6 ↑	78.0 ↑	79.7 ↑	**73.6** ↑
LLaVA-NeXT Liu et al. (2024)	68.3	0.0	70.5	71.3	55.9
+RoboSpatial	**78.9** ↑	19.7 ↑	80.1 ↑	79.0 ↑	70.6 ↑
SpaceLLaVA Chen et al. (2024)	61.0	2.5	61.0	76.2	47.1
+RoboSpatial	71.6 ↑	13.1 ↑	72.4 ↑	**81.8** ↑	67.7 ↑
RoboPoint Yuan et al. (2024)	69.9	19.7	70.5	63.6	44.1
+RoboSpatial	78.0 ↑	**31.1** ↑	**81.0** ↑	70.6 ↑	64.7 ↑
3D VLMs					
3D-LLM Hong et al. (2023)	39.8	0.0	35.2	N/A	N/A
+RoboSpatial	55.2 ↑	8.2 ↑	52.3 ↑	N/A	N/A
LEO Huang et al. (2024b)	51.2	0.0	38.1	N/A	N/A
+RoboSpatial	64.2 ↑	10.0 ↑	57.1 ↑	N/A	N/A
Not available for fine-tuning					
Molmo Deitke et al. (2024)	58.6	0.1	18.1	67.1	55.9
GPT-4o OpenAI et al. (2024)	77.2	5.7	58.1	76.2	70.6

Table 5: Results on an out-of-domain test split comparing prior art VLMs. The results show improved (↑) spatial understanding capabilities on similar domains. Bolded number is the best result for the column.

	Indoor → Tabletop	Tabletop → Indoor
RoboPoint Yuan et al. (2024)	38.7	38.2
+RoboSpatial	48.9 ↑	51.3 ↑
LEO Huang et al. (2024b)	41.9	43.7
+RoboSpatial	47.2 ↑	54.5 ↑

Table 6: Average accuracy for dataset generalization when fine-tuning on indoor scenes and testing on tabletop scenes (indoor→tabletop), and vice versa (tabletop→indoor), evaluated on the RoboSpatial-Val split.

E.4 Accuracy Per Frame of Reference

We show the results per frame in Table 8 for our out-of-domain test set. 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.

E.5 Robot Experiments

We present additional results from our robot experiments in Figure 4 and Figure 5. 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,

Table 7: Results of scaling experiment on LLaVa-Next Liu et al. (2024) with varied spatial relationship annotations. Average accuracy on held-out test set is reported.

Annotation Size	100K	300K	900k (Default)	1.8M	3M (Full)
LLaVa-Next Liu et al. (2024)	38.1	46.7	60.5	65.8	72.4

Table 8: Results of per frame accuracy of existing 2D/3D VLMs on a held-out test split of images and scans. All methods, for all tasks, perform better (↑) when fine-tuned on our ROBOSPATIAL dataset. The best result for each column is bolded.

Model	Indoor			Tabletop			Average		
	Ego-centric	Object-centric	World-centric	Ego-centric	Object-centric	World-centric	Indoor	Tabletop	Total
Open-source VLMs									
2D VLMs									
VILA Lin et al. (2024)	55.9	40.5	32.9	43.6	39.7	28.9	43.1	37.4	40.2
+ROBOSPATIAL	74.3↑	57.8 ↑	62.3 ↑	70.3 ↑	58.1 ↑	60.3 ↑	64.8 ↑	62.9 ↑	63.9 ↑
LLaVA-Next Liu et al. (2024)	35.2	24.3	34.7	36.4	28.5	22.7	31.4	29.2	30.3
+ROBOSPATIAL	75.4 ↑	54.1 ↑	68.8 ↑	67.9 ↑	54.7 ↑	58.9 ↑	60.4 ↑	60.5 ↑	60.5 ↑
SpaceLLaVA Chen et al. (2024)	40.6	36.0	30.1	52.3	32.8	53.5	38.9	46.2	43.6
+ROBOSPATIAL	**78.5** ↑	**60.6** ↑	64.3 ↑	73.0 ↑	49.5 ↑	68.3 ↑	67.8 ↑	63.6 ↑	65.7 ↑
RoboPoint Yuan et al. (2024)	41.9	36.2	40.7	46.2	30.5	37.9	39.6	38.2	38.9
+ROBOSPATIAL	76.4 ↑	58.3 ↑	78.3 ↑	**76.7** ↑	62.6 ↑	71.0 ↑	71.0 ↑	70.1 ↑	70.6 ↑
3D VLMs									
3D-LLM Hong et al. (2023)	28.9	38.3	45.6	38.9	35.7	52.6	37.6	42.4	40.0
+ROBOSPATIAL	60.7 ↑	52.1 ↑	76.5 ↑	57.9 ↑	**62.8** ↑	77.3 ↑	63.1 ↑	66.0 ↑	64.6 ↑
LEO Huang et al. (2024b)	46.9	30.6	48.2	41.4	34.3	55.4	41.9	43.7	42.8
+ROBOSPATIAL	68.1 ↑	71.6 ↑	**79.6** ↑	71.4 ↑	60.2 ↑	**80.5** ↑	**73.1** ↑	**70.7** ↑	**71.9** ↑
Not available for fine-tuning									
2D VLMs									
Molmo Deitke et al. (2024)	50.4	50.8	47.6	64.4	33.6	53.8	49.6	50.6	50.1
GPT-4o OpenAI et al. (2024)	52.9	38.7	56.3	62.5	30.7	63.7	49.3	52.3	50.8

especially with RoboPoint Yuan et al. (2024). We hypothesize that mixing the dataset with RoboPoint training data and ROBOSPATIAL 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.

E.6 MORE QUALITATIVE EXAMPLES

Figure 6 presents 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 OpenAI et al. (2024). 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 Fu et al. (2024), including those involving distance, such as "touching."

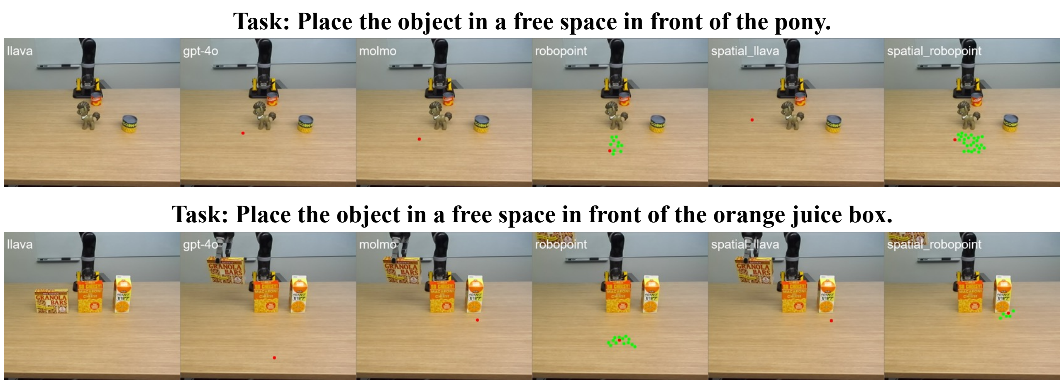


Figure 4: Robotics experiments: the red dot shows the model output (if not present, the model failed to provide a valid point in the image); green dots are used to show when a model outputs multiple points. The robot motion generator, cuRobo Sundaralingam et al. (2023), is used to grasp the item referenced by the generated point. The *spatial-* prefix indicates model trained with RoboSpatial.

Question: pick lone object	
LLaVa-Next Liu et al. (2024)	×
LLaVa-Next-FT Liu et al. (2024)	✓
RoboPoint Yuan et al. (2024)	×
RoboPoint-FT Yuan et al. (2024)	✓
Molmo Deitke et al. (2024)	✓
GPT-4o OpenAI et al. (2024)	×

Question: Is there room to slot the pancake mix in the middle of the row of boxes	
LLaVa-Next Liu et al. (2024)	✓
LLaVa-Next-FT Liu et al. (2024)	✓
RoboPoint Yuan et al. (2024)	×
RoboPoint-FT Yuan et al. (2024)	✓
Molmo Deitke et al. (2024)	✓
GPT-4o OpenAI et al. (2024)	✓

Question: Is there space in the white container for the orange juice box	
LLaVa-Next Liu et al. (2024)	×
LLaVa-Next-FT Liu et al. (2024)	✓
RoboPoint Yuan et al. (2024)	×
RoboPoint-FT Yuan et al. (2024)	×
Molmo Deitke et al. (2024)	×
GPT-4o OpenAI et al. (2024)	✓

Question: alphabet soup fit in the purple box	
LLaVa-Next Liu et al. (2024)	✓
LLaVa-Next-FT Liu et al. (2024)	×
RoboPoint Yuan et al. (2024)	✓
RoboPoint-FT Yuan et al. (2024)	✓
Molmo Deitke et al. (2024)	×
GPT-4o OpenAI et al. (2024)	✓

Question: pick object behind the middle container	
LLaVa-Next Liu et al. (2024)	×
LLaVa-Next-FT Liu et al. (2024)	✓
RoboPoint Yuan et al. (2024)	✓
RoboPoint-FT Yuan et al. (2024)	×
Molmo Deitke et al. (2024)	×
GPT-4o OpenAI et al. (2024)	×

Question: pick shortest object	
LLaVa-Next Liu et al. (2024)	×
LLaVa-Next-FT Liu et al. (2024)	✓
RoboPoint Yuan et al. (2024)	✓
RoboPoint-FT Yuan et al. (2024)	✓
Molmo Deitke et al. (2024)	✓
GPT-4o OpenAI et al. (2024)	✓

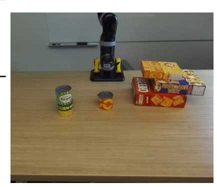
Question: place object in container behind popcorn	
LLaVa-Next Liu et al. (2024)	×
LLaVa-Next-FT Liu et al. (2024)	✓
RoboPoint Yuan et al. (2024)	✓
RoboPoint-FT Yuan et al. (2024)	✓
Molmo Deitke et al. (2024)	×
GPT-4o OpenAI et al. (2024)	×

Question: place the object inside the smallest box	
LLaVa-Next Liu et al. (2024)	×
LLaVa-Next-FT Liu et al. (2024)	✓
RoboPoint Yuan et al. (2024)	✓
RoboPoint-FT Yuan et al. (2024)	✓
Molmo Deitke et al. (2024)	✓
GPT-4o OpenAI et al. (2024)	×

Question: can the robot directly pick the red orange peaches can without disturbing other objects?	
LLaVa-Next Liu et al. (2024)	✓
LLaVa-Next-FT Liu et al. (2024)	✓
RoboPoint Yuan et al. (2024)	×
RoboPoint-FT Yuan et al. (2024)	×
Molmo Deitke et al. (2024)	✓
GPT-4o OpenAI et al. (2024)	✓

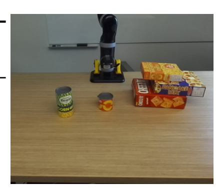
Question: is there an object that is not in a stack?	
LLaVa-Next Liu et al. (2024)	✓
LLaVa-Next-FT Liu et al. (2024)	✓
RoboPoint Yuan et al. (2024)	✓
RoboPoint-FT Yuan et al. (2024)	✓
Molmo Deitke et al. (2024)	✓
GPT-4o OpenAI et al. (2024)	✓

Question: can the macaroni and cheese be placed on top of cheez-it without touching other objects?	
LLaVa-Next Liu et al. (2024)	×
LLaVa-Next-FT Liu et al. (2024)	×
RoboPoint Yuan et al. (2024)	✓
RoboPoint-FT Yuan et al. (2024)	✓
Molmo Deitke et al. (2024)	×
GPT-4o OpenAI et al. (2024)	✓



Question: is there space to place one of the cans on the cheez-it box?	
LLaVa-Next Liu et al. (2024)	×
LLaVa-Next-FT Liu et al. (2024)	×
RoboPoint Yuan et al. (2024)	×
RoboPoint-FT Yuan et al. (2024)	×
Molmo Deitke et al. (2024)	×
GPT-4o OpenAI et al. (2024)	×

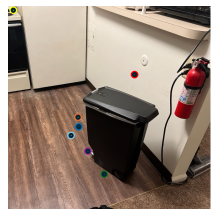
Question: place on the object to the left of macaroni and cheese	
LLaVa-Next Liu et al. (2024)	×
LLaVa-Next-FT Liu et al. (2024)	✓
RoboPoint Yuan et al. (2024)	✓
RoboPoint-FT Yuan et al. (2024)	✓
Molmo Deitke et al. (2024)	✓
GPT-4o OpenAI et al. (2024)	×



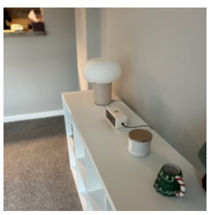
Question: pick the highest object on the stack of two objects	
LLaVa-Next Liu et al. (2024)	×
LLaVa-Next-FT Liu et al. (2024)	×
RoboPoint Yuan et al. (2024)	×
RoboPoint-FT Yuan et al. (2024)	×
Molmo Deitke et al. (2024)	×
GPT-4o OpenAI et al. (2024)	×

Figure 5: 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 Liu et al. (2024)	
LLaVa-Next-FT Liu et al. (2024)	
RoboPoint Yuan et al. (2024)	
RoboPoint-FT Yuan et al. (2024)	
Molmo Deitke et al. (2024)	
GPT-4o OpenAI et al. (2024)	



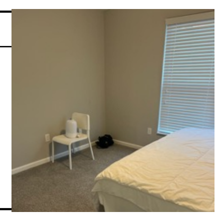
Question: Pinpoint several points within the vacant space situated behind the trash bin.	
Answer	
LLaVa-Next Liu et al. (2024)	
LLaVa-Next-FT Liu et al. (2024)	
RoboPoint Yuan et al. (2024)	
RoboPoint-FT Yuan et al. (2024)	
Molmo Deitke et al. (2024)	
GPT-4o OpenAI et al. (2024)	



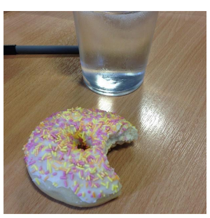
Question: Can the lamp fit in front of the shelf?	
Answer	Yes
LLaVa-Next Liu et al. (2024)	×
LLaVa-Next-FT Liu et al. (2024)	✓
RoboPoint Yuan et al. (2024)	×
RoboPoint-FT Yuan et al. (2024)	✓
Molmo Deitke et al. (2024)	×
GPT-4o OpenAI et al. (2024)	×

Question: Can the pot fit above the fridge?	
Answer	Yes
LLaVa-Next Liu et al. (2024)	×
LLaVa-Next-FT Liu et al. (2024)	✓
RoboPoint Yuan et al. (2024)	×
RoboPoint-FT Yuan et al. (2024)	✓
Molmo Deitke et al. (2024)	×
GPT-4o OpenAI et al. (2024)	×

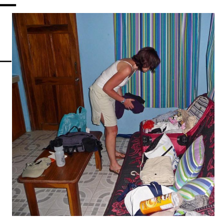
Question: Is the lamp above the shelf?	
Answer	Yes
LLaVa-Next Liu et al. (2024)	×
LLaVa-Next-FT Liu et al. (2024)	✓
RoboPoint Yuan et al. (2024)	×
RoboPoint-FT Yuan et al. (2024)	✓
Molmo Deitke et al. (2024)	×
GPT-4o OpenAI et al. (2024)	✓



Question: Is the chair behind the bed?	
Answer	Yes
LLaVa-Next Liu et al. (2024)	×
LLaVa-Next-FT Liu et al. (2024)	✓
RoboPoint Yuan et al. (2024)	×
RoboPoint-FT Yuan et al. (2024)	✓
Molmo Deitke et al. (2024)	×
GPT-4o OpenAI et al. (2024)	×



Question: Is the dining table touching the donut?	
Answer	Yes
LLaVa-Next Liu et al. (2024)	×
LLaVa-Next-FT Liu et al. (2024)	✓
RoboPoint Yuan et al. (2024)	×
RoboPoint-FT Yuan et al. (2024)	✓
Molmo Deitke et al. (2024)	×
GPT-4o OpenAI et al. (2024)	×



Question: Is the couch under the suit-case?	
Answer	Yes
LLaVa-Next Liu et al. (2024)	×
LLaVa-Next-FT Liu et al. (2024)	✓
RoboPoint Yuan et al. (2024)	×
RoboPoint-FT Yuan et al. (2024)	✓
Molmo Deitke et al. (2024)	×
GPT-4o OpenAI et al. (2024)	×

Figure 6: 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 Fu et al. (2024) dataset, demonstrating that the ROBOSPATIAL-trained model can transfer to unseen relationships.