SpatialBot: Precise Spatial Understanding with Vision Language Models

Wenxiao Cai[1,2,3,*], Iaroslav Ponomarenko[4,*], Jianhao Yuan[5], Xiaoqi Li[4], Wankou Yang[6], Hao Dong[4], Bo Zhao[1,3,†]

***Abstract*— Vision Language Models (VLMs) have achieved impressive performance in 2D image understanding; however, they still struggle with spatial understanding, which is fundamental to embodied AI. In this paper, we propose *SpatialBot*, a model designed to enhance spatial understanding by utilizing both RGB and depth images. To train VLMs for depth perception, we introduce the *SpatialQA* and *SpatialQA-E* datasets, which include multi-level depth-related questions spanning various scenarios and embodiment tasks. *SpatialBench* is also developed to comprehensively evaluate VLMs' spatial understanding capabilities across different levels. Extensive experiments on our spatial-understanding benchmark, general VLM benchmarks, and embodied AI tasks demonstrate the remarkable improvements offered by *SpatialBot*. The model, code, and datasets are available at `https://github.com/BAAI-DCAI/SpatialBot`.**

I. Introduction

Recently, Vision Language Models (VLM) [1], [2], [3], [4], [5], [6] have demonstrated notable capabilities in general 2D visual understanding and reasoning, based on vision encoder-based perception and language model-based reasoning. However, it is still challenging for VLMs to comprehend spatial information from 2D images merely, which is the key to implementing various real-world tasks [7], [8], [9], [10], [11], particularly those embodied AI related tasks such as manipulation [12], [13], [14], [15] and navigation [16], [17], [18], [19].

The main challenges for VLMs to have spatial understanding ability are in the following aspects: 1) Popular VLMs have limited capacity to understand depth information as they are only trained on RGB images without seeing depth images. In addition, the training tasks need little depth information to solve. Consequently, directly inputting depth maps into VLMs results in poor performance. 2) A well-designed dataset for training VLMs to understand depth is absent. The popular VLM tuning datasets provide neither depth maps nor depth-related tasks. 3) The inconsistency of the scales between indoor and outdoor numerical depth is also an important problem preventing VLM from uniformly processing depth in various tasks. For example, tasks such as indoor navigation and manipulation require millimeter-level precision, whereas outdoor tasks do not necessitate such high precision but demand a broader depth range.

[1] Shanghai Jiao Tong University, [2] Stanford University, [3] BAAI, [4] Peking University, [5] University of Oxford, [6] Southeast University

*Wenxiao Cai and Iaroslav Ponomarenko contributed equally to this paper.

†Corresponding author: bo.zhao@sjtu.edu.cn.

*This work was done when Wenxiao Cai was a visiting student in SJTU, and when he was an intern in BAAI.

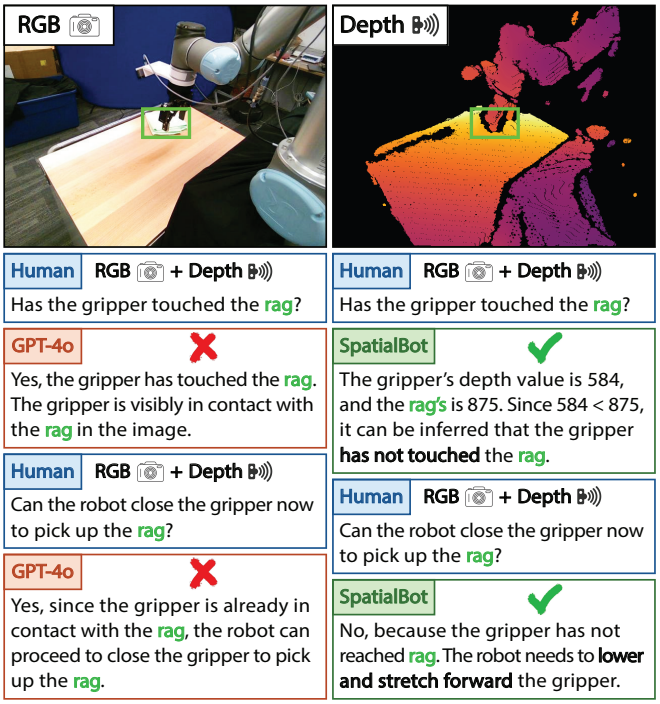

Fig. 1. *SpatialBot* has better spatial understanding ability than GPT-4o. From a single RGB image, even a human can not judge whether the gripper has touched the rag. GPT-4o fails even if looking at both RGB and depth images. *SpatialBot* first obtains depth information of target objects from the depth map, and then judges that the gripper has not yet touched the rag.

To address these challenges, we propose *SpatialBot*, which can precisely comprehend spatial information through depth images and perform robotic manipulation. We design a progressive training approach to first improve the general spatial understanding capacity of VLMs with the proposed *SpatialQA* dataset, which contains general conversation tasks. We then leverage this spatial understanding capacity for embodied tasks using the collected robot manipulation task dataset, *SpatialQA-E*. We design various purpose-specific QA tasks that heavily rely on spatial understanding from low to high levels. These tasks include low-level depth estimation, middle-level object detection, referring QA, and depth comparison, high-level tasks that require depth reasoning, such as understanding spatial relationships in both general conversations and robot manipulation. To enable the model to accurately obtain depth information, we designed a depth API that allows the model to query the depth values of individual pixels or regions.

We validate the spatial comprehension capacity of VLMs with *SpatialBench* which consists of manually annotated question-answer pairs on spatial understanding and reason-

ing. We also deploy *SpatialBot* on robots to do manipulation tasks, for example, picking up the teacup in the middle and placing it on the closest board, as shown in Fig. 4. The experimental results verify that our SpatialBot can understand the depth in the three levels. Furthermore, it is also verified that the fine-tuning of VLMs in SpatialQA can improve their performance on general VLM benchmarks such as MME [20], MMBench [21], etc. Finally, robot manipulation abilities demonstrate the promising applications of *SpatialBot*. In summary, the main contributions of our work are as follows:

- We propose *SpatialBot* that shows promising performance in general visual recognition, spatial understanding, and robot manipulation.
- We curate a large-scale RGB-D VQA dataset, *SpatialQA*, for training *SpatialBot*, and *SpatialBench* for evaluating VLMs' spatial understanding performances. Three levels of tasks have been designed for a comprehensive analysis of depth.
- We finetune and deploy *SpatialBot* on embodiment tasks that involve spatial reasoning, and release the robot manipulation dataset focusing on spatial relationships, namely *SpatialQA-E*.

II. Related work

A. VLM and RGB Datasets

In recent years, VLMs (or Multi-modal Large Language Models (MLLMs)) have achieved significant advancements [22]. LLaVA [3] pioneered visual instruction tuning, which is followed by subsequent works [5], [23], [6], [24], [25] with more extensive datasets [26] and different large language model (LLM) backbones [27], [28], [29], [30]. These VLMs are mainly used to tackle tasks related to perception [20], reasoning [21], and optical character recognition (OCR) [31], [32]. Additionally, some works have introduced an encoder-decoder structure beyond VLMs to perform pixel-level grounding tasks [33], [34], [35], [36], [37], [38]. However, their performances in spatial relationship understanding [39] are mediocre. We posit that comprehending the entire space from a monocular RGB image is overwhelming for VLMs. Integrating depth information could effectively enhance the spatial understanding capabilities of VLMs.

B. Spatial Understanding in General QA and Embodiment

Spatial understanding requires VLMs to understand scenes beyond 2D RGB images. This is particularly crucial in precision tasks such as robotic grasping [40]. Spatial understanding can be achieved through point clouds [40], [41] or depth maps [39]. Some studies have attempted to perform depth estimation [42] and 3D detection [43] directly from monocular RGB images, but the accuracy is limited regarding metric depth estimation. SpatialVLM [44] and SpatialRGPT [45] infer spatial relationships only from 2D images. However, in robotic tasks (see, e.g., Fig. 1), depth information from sensors is essential for spatial understanding. Recently, Monocular Depth Estimation (MDE) has seen rapid advancements. Using large amounts of unsupervised data [46], [47] and synthetic data [48], MDE can accurately estimate the depth in various scenarios [49]. Therefore, we improve the spatial understanding of VLMs by adding depth information to the RGB images they use, leveraging MDE. Despite the strength of monocular depth estimation models, training large models to estimate depth directly is not always feasible. In embodied AI scenarios, precise depth information is required from hardware devices, which depth estimation models cannot achieve. Additionally, enabling VLMs to precisely understand space from a single RGB image has proven to be extremely difficult [42], [43]. To extend spatial understanding abilities to embodiment, we propose *spatialQA-E*. To the best of our knowledge, it is the first manipulation dataset that focuses on spatial relationships. *SpatialBot* utilizes a similar model structure with state-of-the-art vision-language-action models like RT [50], [51], Octo [52] and OpenVLA [53], while acquires spatial knowledge necessary in manipulation tasks through training on *SpatialQA-E*.

III. SpatialBot

We use depth information to guide VLMs in understanding spatial relationships [54], [55]. Compared to point clouds, depth information is easier to collect and process. Moreover, since RGB-D cameras are affordable, most robots carry such cameras to instantly capture RGB and depth images. In addition, due to the remarkable capacities of Monocular Depth Estimation (MDE), one can quickly adapt large-scale RGB datasets to RGB-D datasets in an affordable way.

A. Depth Map Encoding

Our depth encoding aims to preserve all depth information for VLMs to use. A challenge is the indoor and outdoor consistency. Indoor scenes like robot manipulation [13] and indoor navigation [56], [57] may require millimeter-level precision, while outdoor scenes include a large range of depth values. Existing methods often adopt ordinal encoding [58], [46], which, however, cannot be subjected to basic mathematical operations. To address the issue, we use uint24 or three-channel uint8 to store depth values, measured in millimeters from $1mm$ to $131.071m$. We directly save the raw depth values and leave subsequent computations to the powerful fitting capabilities of VLMs. For single-channel uint24, we use millimeters as a unit directly. This way, VLMs can directly query the required values from the depth map. For three-channel uint8 images, we distribute the values across a broader range: the units for the three channels are 2^0, 2^5, and 2^{10} millimeters, respectively. Each channel has 2^5, 2^5, and 2^7 possible values. For an image of size (H,W), to store depth value $d_{H,W}$ (in millimeters) in three-channel uint8 image $I^3_{H,W}$, we encode the image I following:

$$I^0_{h,w} = (d_{h,w}//2^{10}) * 2^1, I^1_{h,w} = (d_{h,w}//2^5) * 2^3, \quad (1)$$

$$I^2_{h,w} = (d_{h,w}\%2^5) * 2^3. \quad (2)$$

The choice of 2^{10} mm as a unit for the first channel is influenced by the depth range in many desktop grasping

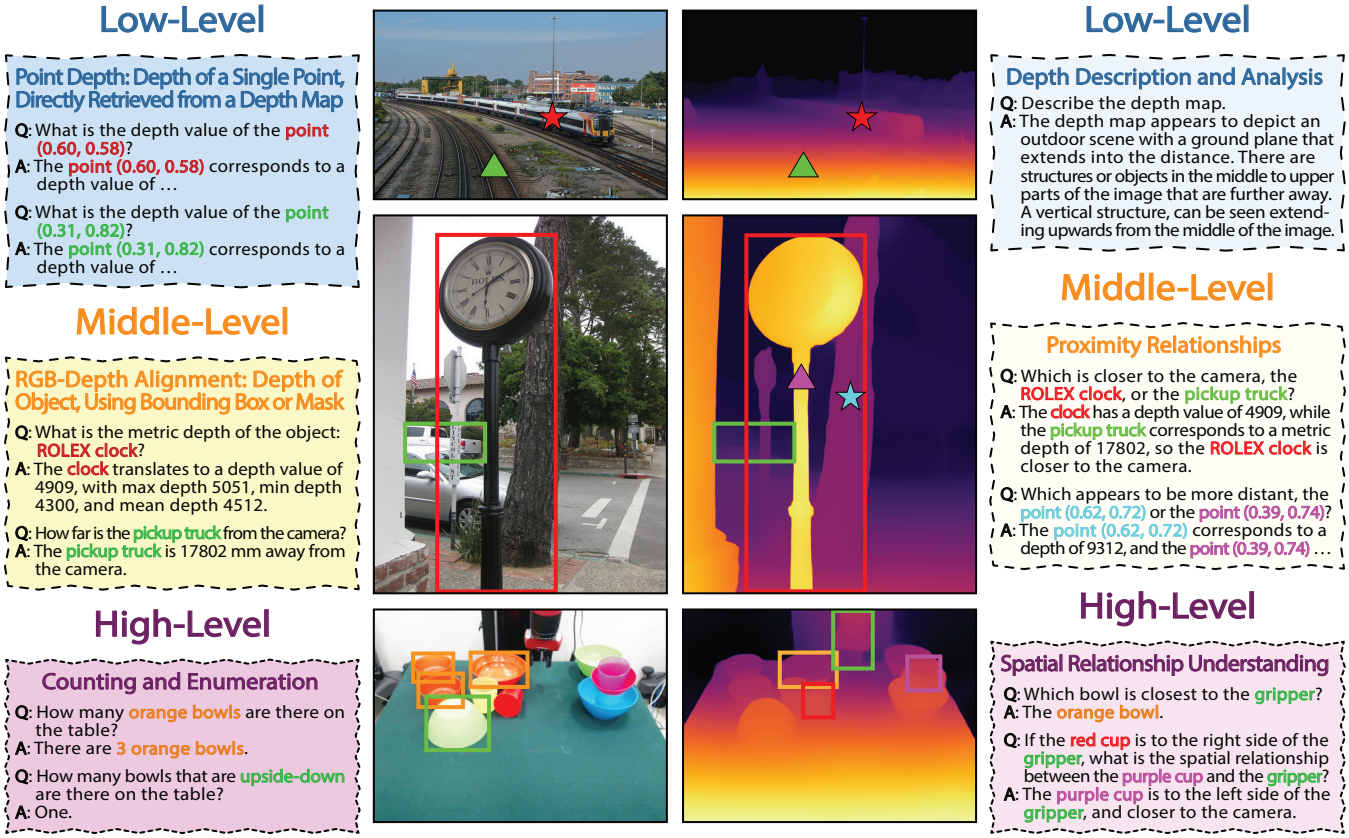

Fig. 2. The proposed *SpatialQA* dataset consists of basic, middle, and high-level VQAs in general VLM tasks, aiming to (a) help VLMs understand depth images, (b) let VLMs learn to align RGB and depth images, (c) enable VLMs to do high-level tasks better by understanding both RGB and depth images, as depth images provide clear boundary information and spatial relationships.

tasks in robotics [59], [60], [61], [62], [63], [64], [65], which typically have a maximum depth of around $1m$. A larger unit would result in the first channel being predominantly zero in most scenarios. Similarly, we use multipliers of 2 and 8 to better distinguish three-channel depth maps. Our experiments have validated that VLMs can easily learn the relationship between our encoding method and the actual depth values.

B. Depth Description of an Object

We do not use a separate image encoder to maintain generality, so *SpatialBot* cannot output pixel-level information. Intuitively, the center point of objects can simply represent their depth. However, for example, in the case of a cup, there is a significant difference between the inner and outer surfaces, so a single value cannot accurately represent the depth. Therefore, we use four depth values—max, min, mean, and center—to describe the object's depth, if its mask is available. Considering that the mask and depth map cannot be precise, we use the 95th and 5th percentile values as the max and min depth values. If accurate bounding boxes are available, we prompt SAM [66] with the bounding box and its central point to get object masks. If not, we use the depth value of the center point of the bounding box.

C. SpatialQA Pipeline

To help VLMs understand depth inputs and use depth information to do high-level tasks like spatial relationship understanding, counting, and enumeration, we design a three-step QA pipeline: (a) This pipeline progressively lets VLM learn to understand depth, align depth and RGB, and use depth for complex reasoning in high-level tasks. (b) Existing RGB datasets can be easily converted to RGB-Depth datasets with our pipeline.

1) Low Level: To enable VLMs to understand depth images and learn to query information from them, we ask for the depth value of points. VLMs should learn to take the depth value directly from depth inputs and relate point coordinates with pixels in the image. Meanwhile, since the visual encoder does not see depth images in pre-training, we also expect the encoder and projector to learn to encode depth images together with RGB images. We also let *SpatialBot* describe the depth map and infer what may be in the images, giving only a depth map.

2) Middle Level: As VLMs have learned to encode and query information from depth images, they should now learn to use depth information. Also, since image and depth inputs are given to VLMs, they should know the relationships between them. First, we ask about proximity relationships, namely, which point is closer or further away. Second, we let VLMs learn to describe the depth of objects or regions by using center point depth, minimum, maximum, and mean depth. VLMs should learn to locate an object in the RGB image and then find depth information from depth input. Third, we ask about proximity relationships between objects.

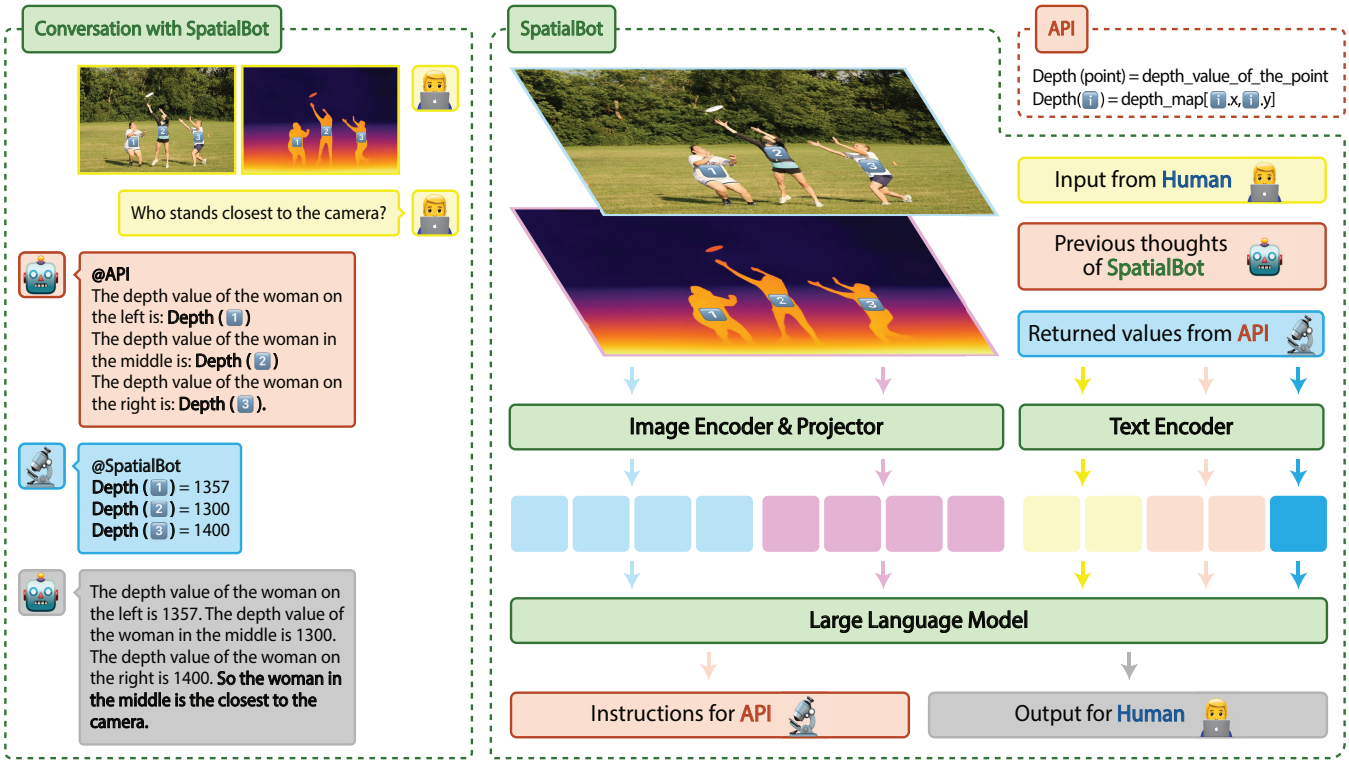

Fig. 3. The architecture of *SpatialBot*. It takes a pair of RGB and depth images as input, where depth images are optional. *SpatialBot* can choose to call Depth API if it needs accurate depth information.

3) High Level: Since VLMs can now understand depth input, align depth with RGB, and have some knowledge about proximity in the spatial world, we design tasks to help VLMs apply depth at a higher level. When the model sees the depth map, the boundaries of objects and their surroundings become clearer, so we believe the depth map aids in grounding and counting tasks. Additionally, in *SpatialQA*, the model develops a clear understanding of space, which aids in determining spatial and positional relationships.

D. SpatialQA-E

We propose *SpatialQA-E* to extend spatial understanding and reasoning abilities to embodiment tasks. We use the 7-axis Franka Research 3 Robotic Arm to grasp objects on the table, avoid obstacles while moving, and place them on a cutting board on the table. We include spatial relationships in language instructions, so the model should learn spatial reasoning in manipulation. *SpatialQA-E* contains 2000 episodes in total. The dataset is composed of 4 steps, shown in Fig. 4 and Fig. 5:

- Learn to pick and place teacups, balls, bananas, etc.
- Find specific object and destination. The dataset includes spatial relationships in positive, comparative (-er), and superlative (-est) degrees from the perspective of the robot or the human (camera):
  - Positional: left/ right/ middle/ up/ down on/ in/ inside/ outside
  - Size: tall/ short/ large/ small/ wide/ thin/ big/ small
  - Illusion: we take photos of objects, print them out, and put the printed object on the table. It looks real, and the model needs to tell between printed and real objects through visual clues, e.g., depth information (printed objects are flat) and shadows.
- In moving objects, the robot needs to avoid obstacles.

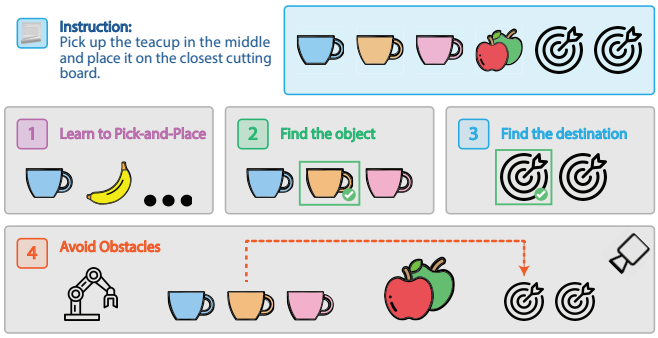

Fig. 4. *SpatialQA-E* involves spatial relationships in robot manipulation.

E. SpatialBot in Embodiment Tasks

SpatialBot is finetuned on *SpatialQA-E* to work on embodiment tasks. In short, it is a Vision-Language-Action (VLA) model that supports multi-frame RGB or RGB-D inputs. Robot manipulation tasks are specified as: in current time stamp t, given history and current image observations $x_{j=0}^{t}$ (RGB or RGB-D), models should learn policy $\pi(|i, x_{j=0}^{t})$. Action a_t is sampled from π and applied to robots. For a robot of two-finger end effector, action space can be represented as 7 DoF vector: $(\Delta X, \Delta Y, \Delta Z, \Delta R, \Delta P, \Delta Yaw, C)$, indicating delta change in poses XYZ and rotation $RPYa$ (roll, pitch, yaw), gripper closure C. The delta change in position and rotation of action space is encoded into 101 possible values, from 0, 0.01 to 1. The model output texts of

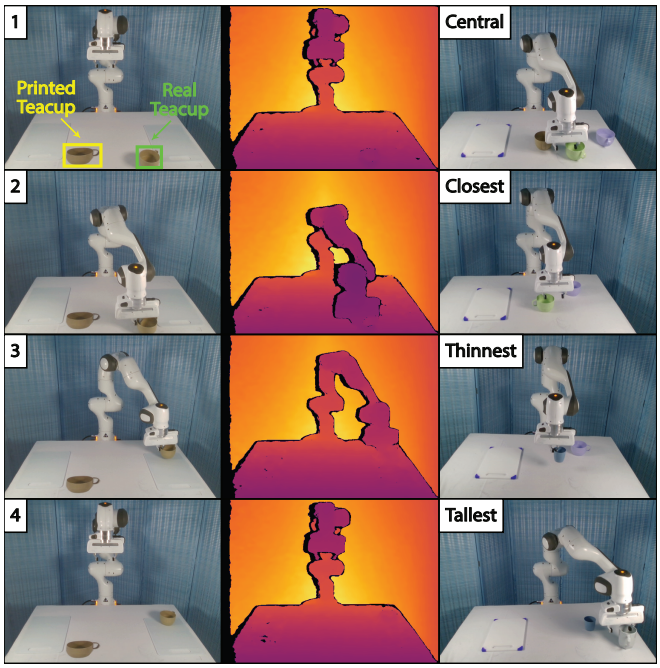

Fig. 5. *SpatialQA-E* demonstration. Left: 4 steps in picking up the real teacup and putting it on the right cutting board relative to the camera. We print the teacup as a distraction. It's easier to tell between the real and printed teacup from the depth map. Right: 4 sample settings in *SpatialQA-E*, where we specify spatial relationships.

7 DoF actions directly. A sample conversation: 'User: What should the robot do to pick up the biggest teacup and move it to the left cutting board? Answer with robot parameters. - SpatialBot: The robot should <0.17, 0.51, 0.44, 0.62, 0.83, 0.07, 1>'. Then we decode the output to robot control signals to control the robot movement of each frame. If the model directly answers robotic parameters during the finetuning stage, we find that it can only respond to robot-specific questions. To enable multi-task training, we incorporate some natural language elements into the robot's responses, such as 'The robot should'. Then, we train the model on robotic data and general QA data, such as SpatialQA-E and SpatialQA. We have the model predict special tokens during robotic tasks to maintain the model's numerical reasoning abilities in general conversations. We predict each frame's delta pose instead of the target pose. This choice allows for more precise control of the robot by dividing each dimension of the action space into 100 bins.

F. SpatialBench

To evaluate VLM's performance on general tasks, we annotate the *SpatialBench*. The following questions are included:

- Has [Object 1] touched/reached [Object 2]?
- What is the spatial relationship between [Object]s?
- Counting and enumeration.
- Size comparison between objects.

G. SpatialBot Model with Depth API

To enable the model to accurately obtain depth information, we design **Depth API**. When the *SpatialBot*'s output contains text with a format of Depth(point), the API will query the depth value of that point in the corresponding depth map and then input this depth value back into *SpatialBot*. Combining the user's question with the API's return value, *SpatialBot* will provide the final answer. The model can call the API to get the precise depth value of a specific point. For example, when *SpatialBot* wants to know the depth information of an object, it first determines the bounding box of the object and then calls the Depth API using the center point of the bounding box. If the model wants to obtain the depth range of this object, it first observes which points in the image correspond to the maximum and minimum depth values and calls the Depth API using the coordinates of these points. However, to enhance the model's understanding of the depth map itself, we only allow *SpatialBot* to call the API on a subset of the data during training.

IV. Experiments

We start with validating that *SpatialBot* can understand depth, extract information from depth maps, and perform high-level tasks in *SpatialBench*. Then, we show model performance in general VLM tasks by introducing depth maps and training on *SpatialQA*. Finally, experiments on embodiment tasks show that *SpatialBot* benefits from understanding depth in pick-and-place tasks.

A. SpatialBench

We compare model performance on our *SpatialBench*, which is composed of positional relationship, object existence, reaching, and size comparison tasks. GPT-4o and LLaVA-v1.6-34B [3] without training on *SpatialQA* are compared with models trained on *SpatialQA*. 3B, 4B, and 8B models trained on *SpatialQA* reaches comparable results with GPT-4o. Results are reported in Table I. We do not report GPT-4o results on depth estimation because they were extremely poor. We're uncertain whether this was due to suboptimal prompts or if GPT-4o simply lacks the capability for this task.

B. General VLM Benchmarks

We report results on general benchmarks: MME perception [20] (MME^P), MME cognition (MME^C), MM-Bench [21] test and dev set (MMB^T and MMB^D), SEED Bench Image [31] (SEED(-I)), VQA [67] test-dev split (VQA^{v2}), GQA [68], and POPE [69] (the averaged F1-score of three categories on the validation set of COCO). In most of these benchmarks, RGB information alone is enough. We only use RGB-Depth input on MME^P and GQA, since they contain counting, existence, and position questions, where we expect depth information can benefit such cases.

C. SpatialBot in Embodiment Tasks

We finetune *SpatialBot* on *SpatialQA-E* to do manipulation tasks on real robots. It can be seen as a VLA model supporting multi-frame RGB or RGBD inputs. We use QWen-1.5-0.5B [30] as the base LLM and CLIP [70] as the vision encoder. The pretrain dataset is Bunny-pretrain-LAION-2M [6], and *SpatialQA-E* is used in finetuning. Four

TABLE I

RESULTS ON *SpatialBench*. THE BEST RESULTS OF MODELS WITH THE SAME BASE LLMS ARE MARKED WITH **BOLD** TEXT. LLM-RGB AND LLM-RGBD ARE TRAINED ON RGB IMAGES ONLY AND TESTED WITH RGB AND RGBD INPUTS, RESPECTIVELY. *SpatialBot* WITH RGB INPUT IN DEPTH ESTIMATION IS THE SAME AS THE MDE TASK.

Model	Depth ↑	Position ↑	Existence ↑	Counting ↑	Reaching ↑	Size ↑
GPT-4o-RGB	-	70.6	85.0	84.5	51.7	43.3
GPT-4o-RGBD	-	61.8	90.0	85.2	51.7	40.0
Bunny-Phi2-3B-RGB	70.6	50.0	75.0	89.4	51.7	26.7
SpatialBot-Phi2-3B-RGB	84.1	**64.7**	**80.0**	88.0	**61.7**	**28.3**
Bunny-Phi2-3B-RGBD	85.8	50.0	75.0	90.4	43.3	**28.3**
SpatialBot-Phi2-3B-RGBD	**>99**	61.8	**80.0**	**91.7**	55.0	26.7
Bunny-Phi3-4B-RGB	32.3	58.8	**75.0**	91.0	31.7	16.7
SpatialBot-Phi3-4B-RGB	83.2	64.7	**75.0**	91.0	**40**	**23.3**
Bunny-Phi3-4B-RGBD	63.3	52.9	60.0	85.4	31.7	18.3
SpatialBot-Phi3-4B-RGBD	**>99**	**67.7**	70.0	**91.7**	35.0	21.7
Bunny-QWen-1.5-4B-RGB	42.2	50.0	**75.0**	**91.6**	26.7	15.0
SpatialBot-QWen1.5-4B-RGB	89.9	**52.9**	**75.0**	88.6	**46.8**	18.3
Bunny-QWen-1.5-4B-RGBD	74.6	44.1	70.0	90.7	25.0	15.0
SpatialBot-QWen1.5-4B-RGBD	**>99**	**52.9**	60.0	90.5	41.7	**26.7**
Bunny-Llama3-8B-RGB	58.1	50.0	75.0	91.7	38.3	23.3
SpatialBot-Llama3-8B-RGB	85.6	**55.9**	**80.0**	**91.2**	40.0	20.0
Bunny-Llama3-8B-RGBD	64.0	50.0	75.0	90.4	38.3	**25.0**
SpatialBot-Llama3-8B-RGBD	**>99**	53.0	75.0	90.4	**45.0**	20.0

TABLE II

RESULTS ON GENERAL VLM BENCHMARKS. FOR THE SAME BASE LLM MODELS, BETTER RESULTS ARE MARKED WITH **BOLD** TEXT. RGB-D INPUTS ARE ONLY USED IN MME. WE REPORT THE RESULTS OF BUNNY TRAINED WITH RGB AND TESTED WITH RGB/RGB-D IN IT, SPLIT WITH SLASH. *SpatialBot* IS TRAINED ON RGBD AND TESTED ON RGB/RGB-D ON MME.

Model	MME^P ↑	MME^C ↑	MMB^T ↑	MMB^D ↑	SEED-I ↑	VQA^{v2} ↑	GQA ↑	POPE ↑
Bunny-Phi2-3B	1472/1474	286/285	67.90	**68.90**	69.91	78.98	61.52	86.21
SpatialBot-Phi2-3B	1483/**1487**	310/**312**	**70.12**	68.56	**70.85**	**79.80**	**62.28**	**87.04**
Bunny-Phi3-4B	1417/1364	308/319	70.15	70.74	71.04	**80.57**	61.18	84.60
SpatialBot-Phi3-4B	**1431**/1433	**337**/329	**73.49**	**73.11**	**71.64**	80.01	**62.16**	**85.47**
Bunny-QWen1.5-4B	1340/1364	251/254	69.56	68.56	70.05	**80.63**	61.55	85.10
SpatialBot-QWen1.5-4B	1378/**1406**	266/**285**	**70.91**	**69.67**	**70.36**	79.69	**62.77**	**86.09**
Bunny-Llama3-8B	1574/1542	342/318	73.67	74.15	72.32	80.50	62.18	85.22
SpatialBot-LLama3-8B	**1577**/1576	**352**/333	**75.78**	**74.83**	**72.40**	**80.94**	**62.90**	**85.33**

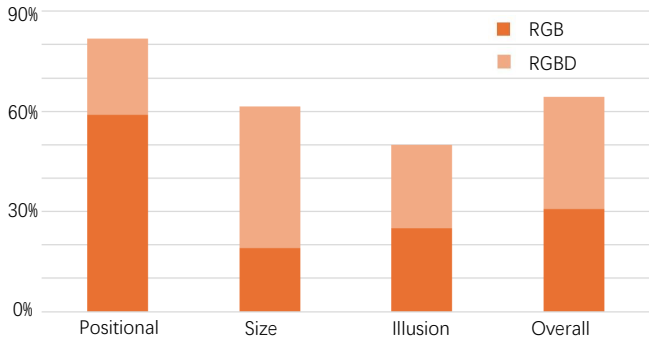

Fig. 6. *SpatialBot* success rate in pick-and-place of RGB and RGBD variants.

frames in history are used to predict the end-effector delta position of the current frame. The model runs locally or connects through an ssh/sftp connection to run on RTX 4090 GPU. It is validated through experiments that *SpatialBot* can do manipulation tasks with spatial instructions. Fig. 6 shows the success rate of *SpatialBot* RGB and RGBD variants. With depth information, *SpatialBot* can pick and place more accurately. **Please refer to our supplementary video for demonstration.**

V. CONCLUSION

We propose *SpatialBot*, a family of state-of-the-art VLMs, for effective depth understanding and thus precise robot manipulating in embodied AI by training on our constructed *SpatialQA* and *SpatialQA-E* datasets. *SpatialBench* is also designed to evaluate the model performance of spatial knowledge in multiple aspects. Experimental results on our benchmark, general VLM benchmarks, and robot manipulation deployment verify the effectiveness and superiority of *SpatialBot* comparing to competitors.

ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China under No. 62306046 and the National Youth Talent Support Program under No. 8200800081.

REFERENCES

[1] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang, "The dawn of lmms: Preliminary explorations with gpt-4v(ision)," *ArXiv*, vol. abs/2309.17421, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263310951

[2] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[3] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *ArXiv*, vol. abs/2304.08485, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258179774

[4] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International Conference on Machine Learning*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256390509

[5] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:261101015

[6] M. He, Y. Liu, B. Wu, J. Yuan, Y. Wang, T. Huang, and B. Zhao, "Efficient multimodal learning from data-centric perspective," *ArXiv*, vol. abs/2402.11530, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267751050

[7] X. Pan, J. Shi, P. Luo, X. Wang, and X. Tang, "Spatial as deep: Spatial cnn for traffic scene understanding," in *AAAI Conference on Artificial Intelligence*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:9164115

[8] D. Azuma, T. Miyanishi, S. Kurita, and M. Kawanabe, "Scanqa: 3d question answering for spatial scene understanding," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19 107–19 117, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:245334889

[9] W. Flores-Fuentes, G. Trujillo-Hernández, I. Y. Alba-Corpus, J. C. Rodríguez-Quiñonez, J. E. Mirada-Vega, D. Hernández-Balbuena, F. N. Murrieta-Rico, and O. Y. Sergiyenko, "3d spatial measurement for model reconstruction: A review," *Measurement*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:255331893

[10] Z. Kang, J. Yang, Z. Yang, and S. Cheng, "A review of techniques for 3d reconstruction of indoor environments," *ISPRS Int. J. Geo Inf.*, vol. 9, p. 330, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219492326

[11] E. Y. Lam, "Computational photography with plenoptic camera and light field capture: tutorial." *Journal of the Optical Society of America. A, Optics, image science, and vision*, vol. 32 11, pp. 2021–32, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:28928392

[12] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," *ArXiv*, vol. abs/2109.12098, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:237396838

[13] A. Padalkar, A. Pooley, A. Jain, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Singh, A. Brohan, *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models," *arXiv preprint arXiv:2310.08864*, 2023.

[14] X. Li, M. Zhang, Y. Geng, H. Geng, Y. Long, Y. Shen, R. Zhang, J. Liu, and H. Dong, "Manipllm: Embodied multimodal large language model for object-centric robotic manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 18 061–18 070.

[15] J. Gao, B. Sarkar, F. Xia, T. Xiao, J. Wu, B. Ichter, A. Majumdar, and D. Sadigh, "Physically grounded vision-language models for robotic manipulation," *arXiv preprint arXiv:2309.02561*, 2023.

[16] S. Levine and D. Shah, "Learning robotic navigation from experience: principles, methods and recent results," *Philosophical Transactions of the Royal Society B*, vol. 378, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:254564889

[17] S. Werner, B. Krieg-Brückner, H. A. Mallot, K. Schweizer, and C. Freksa, "Spatial cognition: The role of landmark, route, and survey knowledge in human and robot navigation," in *GI Jahrestagung*, 1997. [Online]. Available: https://api.semanticscholar.org/CorpusID:16607240

[18] J. Crespo, J. C. Castillo, Ó. M. Mozos, and R. Barber, "Semantic information for robot navigation: A survey," *Applied Sciences*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:214429751

[19] J. Yuan, S. Sun, D. Omeiza, B. Zhao, P. Newman, L. Kunze, and M. Gadd, "Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model," *arXiv preprint arXiv:2402.10828*, 2024.

[20] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, Z. Qiu, W. Lin, J. Yang, X. Zheng, K. Li, X. Sun, and R. Ji, "Mme: A comprehensive evaluation benchmark for multimodal large language models," *ArXiv*, vol. abs/2306.13394, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259243928

[21] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, K. Chen, and D. Lin, "Mmbench: Is your multi-modal model an all-around player?" *ArXiv*, vol. abs/2307.06281, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259837088

[22] Y. Jin, J. Li, Y. Liu, T. Gu, K. Wu, Z. Jiang, M. He, B. Zhao, X. Tan, Z. Gan, Y. Wang, C. Wang, and L. Ma, "Efficient multimodal large language models: A survey," *ArXiv*, vol. abs/2405.10739, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269899856

[23] A. Young, B. Chen, C. Li, C. Huang, G. Zhang, G. Zhang, H. Li, J. Zhu, J. Chen, J. Chang, *et al.*, "Yi: Open foundation models by 01. ai," *arXiv preprint arXiv:2403.04652*, 2024.

[24] B. McKinzie, Z. Gan, J.-P. Fauconnier, S. Dodge, B. Zhang, P. Dufter, D. Shah, X. Du, F. Peng, F. Weers, *et al.*, "Mm1: Methods, analysis & insights from multimodal llm pre-training," *arXiv preprint arXiv:2403.09611*, 2024.

[25] Y. Li, Y. Zhang, C. Wang, Z. Zhong, Y. Chen, R. Chu, S. Liu, and J. Jia, "Mini-gemini: Mining the potential of multi-modality vision language models," *ArXiv*, vol. abs/2403.18814, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268724012

[26] B. Zhao, B. Wu, M. He, and T. Huang, "Svit: Scaling up visual instruction tuning," *arXiv preprint arXiv:2307.04087*, 2023.

[27] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," *ArXiv*, vol. abs/2302.13971, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257219404

[28] M. Abdin, S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, H. Behl, *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv preprint arXiv:2404.14219*, 2024.

[29] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," *See https://vicuna. lmsys. org (accessed 14 April 2023)*, vol. 2, no. 3, p. 6, 2023.

[30] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[31] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan, "Seed-bench: Benchmarking multimodal llms with generative comprehension," *ArXiv*, vol. abs/2307.16125, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260334888

[32] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen, "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi," *ArXiv*, vol. abs/2311.16502, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265466525

[33] H. Zhang, H. Li, F. Li, T. Ren, X. Zou, S. Liu, S. Huang, J. Gao, L. Zhang, C. yue Li, and J. Yang, "Llava-grounding: Grounded visual chat with large multimodal models," *ArXiv*, vol. abs/2312.02949, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265659110

[34] H. You, H. Zhang, Z. Gan, X. Du, B. Zhang, Z. Wang, L. Cao, S.-F. Chang, and Y. Yang, "Ferret: Refer and ground anything anywhere at any granularity," *ArXiv*, vol. abs/2310.07704, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263834718

[35] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia, "Lisa: Reasoning segmentation via large language model,"

ArXiv, vol. abs/2308.00692, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260351258

[36] Y. Yuan, W. Li, J. Liu, D. Tang, X. Luo, C. Qin, L. Zhang, and J. Zhu, "Osprey: Pixel understanding with visual instruction tuning," *ArXiv*, vol. abs/2312.10032, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266335219

[37] A. Zhang, W. Ji, and T.-S. Chua, "Next-chat: An lmm for chat, detection and segmentation," *ArXiv*, vol. abs/2311.04498, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265051059

[38] J. Yang, R. Ding, E. L. Brown, X. Qi, and S. Xie, "V-irl: Grounding virtual intelligence in real life," *ArXiv*, vol. abs/2402.03310, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267412337

[39] J. Jain, J. Yang, and H. Shi, "Vcoder: Versatile vision encoders for multimodal large language models," *ArXiv*, vol. abs/2312.14233, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266521081

[40] H. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11 441–11 450, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219964473

[41] Z. Xu, C. Gao, Z. Liu, G. Yang, C. Tie, H. Zheng, H. Zhou, W. Peng, D. Wang, T. Chen, Z. Yu, and L. Shao, "Manifoundation model for general-purpose robotic manipulation of contact synthesis with arbitrary objects and robots," 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269757051

[42] J. Li, X. Nan, M. Lu, L. Du, and S. Zhang, "Proximity qa: Unleashing the power of multi-modal large language models for spatial proximity analysis," *ArXiv*, vol. abs/2401.17862, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267334932

[43] J. H. Cho, B. Ivanovic, Y. Cao, E. Schmerling, Y. Wang, X. Weng, B. Li, Y. You, P. Krahenbuhl, Y. Wang, and M. Pavone, "Language-image models with 3d understanding," 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269605134

[44] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Driess, P. Florence, D. Sadigh, L. Guibas, and F. Xia, "Spatialvlm: Endowing vision-language models with spatial reasoning capabilities," *ArXiv*, vol. abs/2401.12168, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267069344

[45] A.-C. Cheng, H. Yin, Y. Fu, Q. Guo, R. Yang, J. Kautz, X. Wang, and S. Liu, "Spatialrgpt: Grounded spatial reasoning in vision language model," 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270215984

[46] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," *ArXiv*, vol. abs/2401.10891, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267061016

[47] S. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Muller, "Zoedepth: Zero-shot transfer by combining relative and metric depth," *ArXiv*, vol. abs/2302.12288, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257205739

[48] B. W. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," *ArXiv*, vol. abs/2312.02145, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265609019

[49] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 1623–1637, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:195776274

[50] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. J. Joshi, R. C. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K.-H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. S. Ryoo, G. Salazar, P. R. Sanketi, K. Sayed, J. Singh, S. A. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. H. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "Rt-1: Robotics transformer for real-world control at scale," *ArXiv*, vol. abs/2212.06817, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:254591260

[51] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, K. Choromanski, T. Ding, D. Driess, K. A. Dubey, C. Finn, P. R. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. J. Joshi, R. C. Julian, D. Kalashnikov, Y. Kuang, I. Leal, S. Levine, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. S. Ryoo, G. Salazar, P. R. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. H. Vuong, A. Wahid, S. Welker, P. Wohlhart, T. Xiao, T. Yu, and B. Zitkovich, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *ArXiv*, vol. abs/2307.15818, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260293142

[52] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, *et al.*, "Octo: An open-source generalist robot policy," *arXiv preprint arXiv:2405.12213*, 2024.

[53] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[54] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. A. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. J. Davison, and A. W. Fitzgibbon, "Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera," *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:3345516

[55] H. Pan, T. Guan, Y. Luo, L. Duan, Y. Tian, L. Yi, Y. Zhao, and J. Yu, "Dense 3d reconstruction combining depth and rgb information," *Neurocomputing*, vol. 175, pp. 644–651, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:38137926

[56] G. L. Blanc, Y. Mezouar, and P. Martinet, "Indoor navigation of a wheeled mobile robot along visual routes," *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pp. 3354–3359, 2005. [Online]. Available: https://api.semanticscholar.org/CorpusID:17799387

[57] F. Gul, W. Rahiman, and S. S. N. Alhady, "A comprehensive study for robot navigation techniques," *Cogent Engineering*, vol. 6, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:198355741

[58] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2002–2011, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:46968214

[59] L. Y. Chen, S. Adebola, and K. Goldberg, "Berkeley UR5 demonstration dataset," https://sites.google.com/view/berkeley-ur5/home.

[60] H. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, T. Zhao, P. Hansen-Estruch, Q. H. Vuong, A. W. He, V. Myers, K. Fang, C. Finn, and S. Levine, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:261100981

[61] Y. Feng, N. Hansen, Z. Xiong, C. Rajagopalan, and X. Wang, "Finetuning offline world models in the real world," *ArXiv*, vol. abs/2310.16029, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264438898

[62] H. Shi, H. Xu, S. Clarke, Y. Li, and J. Wu, "Robocook: Long-horizon elasto-plastic object manipulation with diverse tools," *ArXiv*, vol. abs/2306.14447, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259251806

[63] Y. Wang, Z. Li, M. Zhang, K. Driggs-Campbell, J. Wu, L. Fei-Fei, and Y. Li, "D3fields: Dynamic 3d descriptor fields for zero-shot generalizable robotic manipulation," *ArXiv*, vol. abs/2309.16118, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263134320

[64] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv preprint arXiv:2403.12945*, 2024.

[65] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. J. Fan, "Vima: General robot manipulation with multimodal prompts," *ArXiv*, vol. abs/2210.03094, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252735175

[66] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. B. Girshick, "Segment anything," *2023 IEEE/CVF*

International Conference on Computer Vision (ICCV), pp. 3992–4003, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257952310

[67] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," *International Journal of Computer Vision*, vol. 127, pp. 398 – 414, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:8081284

[68] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:152282269

[69] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J. rong Wen, "Evaluating object hallucination in large vision-language models," in *Conference on Empirical Methods in Natural Language Processing*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258740697

[70] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:231591445