SoFar: Language-Grounded Orientation Bridges Spatial Reasoning and Object Manipulation

Zekun Qi[13*] **Wenyao Zhang**[237*] **Yufei Ding**[34*] **Runpei Dong**[5] **Xinqiang Yu**[3]
Jingwen Li[4] **Lingyun Xu**[4] **Baoyu Li**[5] **Xialin He**[5] **Guofan Fan**[1] **Jiazhao Zhang**[3] **Jiawei He**[3]
Jiayuan Gu[6] **Xin Jin**[7] **Kaisheng Ma**[1] **Zhizheng Zhang**[3‡] **He Wang**[34‡] **Li Yi**[18‡]

[1]Tsinghua University [2]Shanghai Jiao Tong University [3]Galbot [4]Peking University [5]UIUC
[6]ShanghaiTech University [7]Eastern Institute of Technology [8]Shanghai Qi Zhi Institute

Project Page GitHub Code HuggingFace

Figure 1: We introduce the concept of *Semantic Orientation*, which refers to natural language-grounded object orientations, such as the "cutting" direction of a knife or the "handle" direction of a cup. To support this, we construct OrienText300K, a large-scale object-text-orientation pairs dataset.

Abstract

While spatial reasoning has made progress in object localization relationships, it often overlooks object orientation—a key factor in 6-DoF fine-grained manipulation. Traditional pose representations rely on pre-defined frames or templates, limiting generalization and semantic grounding. In this paper, we introduce the concept of semantic orientation, which defines object orientations using natural language in a reference-frame-free manner (*e.g.*, the "plug-in" direction of a USB or the "handle" direction of a cup). To support this, we construct OrienText300K, a large-scale dataset of 3D objects annotated with semantic orientations, and develop PointSO, a general model for zero-shot semantic orientation prediction. By integrating semantic orientation into VLM agents, our SoFar framework enables 6-DoF spatial reasoning and generates robotic actions. Extensive experiments demonstrated the effectiveness and generalization of our SoFar, *e.g.*, zero-shot 48.7% successful rate on Open6DOR and zero-shot 74.9% successful rate on SIMPLER-Env.

*Equal contribution. ‡Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Figure 2: **Representation comparison between semantic orientation and others**.

1 Introduction

We observe that current VLMs struggle with understanding object **orientation**, making them insufficient for 6-DoF robot manipulation planning. Consider some everyday scenarios: cutting bread in half with a knife, righting a tilted wine glass, or plugging a cord into a power strip. Previous approaches [10, 12, 8] primarily focused on understanding "*where are the knife and wine glass*" while ignoring their orientations—such as the "blade direction" of the knife and the "up direction" of the glass. This oversight makes it challenging to accomplish these 6-DoF manipulation tasks.

More importantly, different orientations of an object hold varying semantic significance. The capability of connecting specific orientations to their semantic meanings is essential for language-guided robot manipulations. For example, inserting a pen into a pen holder requires aligning the pen tip with the direction of the pen holder's opening; righting a wine glass necessitates aligning the glass's top with the z-axis in the world coordinate frame; and plugging into a power strip involves understanding the "insertion" direction, which is perpendicular to the power strip's surface. However, translating a specific language description into a desired orientation is challenging for existing VLMs.

To move forward, we introduce *language-grounded orientation that bridges spatial reasoning and object manipulation*, characterized by the following:

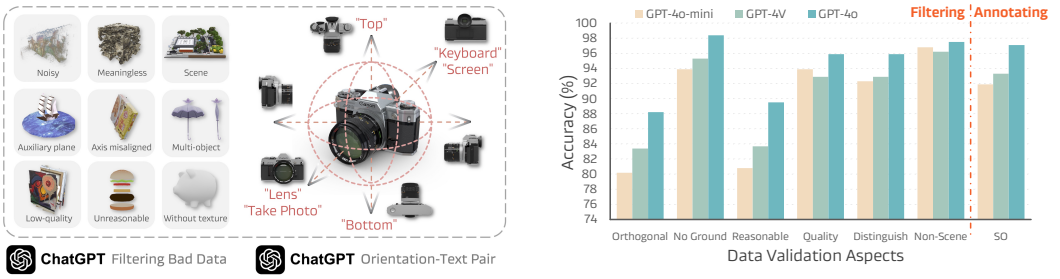
- **From Position Awareness to Orientation Awareness.** While prior works [10, 12, 8] emphasize position relationship, orientation understanding is equally critical for defining the full 6-DoF of object pose or end-effector poses [16, 120, 124, 60]. Orientation awareness involves understanding object orientations and their relationships in the open world, enabling robots to complete tasks requiring precise alignment and rearrangement.

- **From Orientation to Semantic Orientation.** Traditional orientation, defined relative to a base frame or template model [104, 58, 120, 16], is insufficient for open-world manipulation guided by language instructions [108, 49]. We introduce semantic orientation, linking orientational vectors of an object to open-vocabulary prompts (*e.g.*, the "handle" direction of a knife or "plug-in" direction of a USB). This bridges geometric reasoning with functional semantics, enabling robots to interpret task-specific orientation changes.

Achieving such open-world orientation understanding requires rich world knowledge. To this end, we design both the model architecture and the dataset accordingly. We propose **PointSO**, a generalizable cross-modal 3D Transformer [114, 26, 89, 91] for semantic orientation prediction. To train it at scale, we construct **OrienText300K**, a large-scale dataset comprising over 350K 3D models with diverse orientation-text pairs. These annotations are from Objaverse [20] and generated automatically by prompting GPT-4o [48] with rich semantic queries covering both intra-object spatial reasoning and inter-object manipulation contexts—eliminating the need for costly robot-collected data.

To enable comprehensive spatial reasoning, we develop **SoFar**, an integrated system that combines PointSO with foundation models such as SAM [57]. Given an RGB-D input, SAM segments the scene, and PointSO estimates object orientations to build an orientation-aware 3D scene graph. The graph together with the image is fed into a VLM to generate chain-of-thought [119] spatial reasoning, supporting both positional and orientational planning for downstream robotic manipulation.

In addition, we introduce Open6DOR V2, a large-scale benchmark for 6-DoF object rearrangement in simulation, which supports both open-loop and closed-loop control. Our method significantly outperforms state-of-the-art VLMs and VLA models—even those trained on expensive robot trajectories—across both simulated and real-world tasks. We also introduce 6-DoF SpatialBench, a new spatial visual-question-answering benchmark to rigorously assess orientation-aware reasoning.



(a) Data Construction of OrienText300K.

(b) Data filtering and annotating accuracy.

Figure 3: **Visualization of OrienText300K data construction and validation results**.

In summary, we propose **Semantic Orientation** as a new representation that bridges spatial reasoning and robotic manipulation, enabling open-vocabulary, template-free orientation understanding for unseen objects. We introduce **OrienText300K**, a large-scale dataset including 350K diverse objects & orientations and 8M images through careful filtering and annotating. We develop the **SoFar** system, which enhances spatial reasoning with 6-DoF scene graph and achieves SOTA performance on Open6DOR, SimplerEnv, and generalizes across embodiments (*e.g.*, grippers, suction cups, dexterous hands) and tasks (*e.g.*, manipulation, navigation, VQA) without any task-specific fine-tuning. Finally, we present two new benchmarks, **Open6DOR V2** and **6-DoF SpatialBench**, to evaluate 6-DoF rearrangement and spatial reasoning.

2 Semantic Orientation: Connecting Language and Object Orientation

2.1 Definition of Semantic Orientation

Traditionally, object orientation is defined within a reference frame using quaternions or Euler angles to describe relative rotations. However, in interactive tasks, orientations often carry semantic meaning. Humans naturally interpret orientation in a semantic, reference-free manner. For example, plugging in a charger involves aligning the metal prongs with the socket's opening direction—a semantically grounded alignment. Motivated by this, we define an object's *Semantic Orientation* as a unit vector that captures the direction corresponding to a given language description. Formally, for an object X and a description ℓ, the semantic orientation $\mathbf{s}_\ell^X \in S(2)$ is defined as:

$$\mathbf{s}_\ell^X = \mathcal{F}(X, \ell). \tag{1}$$

Here, ℓ is open-vocabulary phrase referring to general directions (*e.g.*, *front*, *top*), object parts (*e.g.*, *handle*, *cap*), or interactions (*e.g.*, *pour out*, *plug-in*). An object X can be associated with multiple semantic orientations by varying the language input, forming a set $S_X = \{\mathbf{s}_{\ell_1}^X, \mathbf{s}_{\ell_2}^X, \ldots, \mathbf{s}_{\ell_n}^X\}$. These orientations provide a semantic basis for describing and transforming the object's rotation.

2.2 OrienText300K: Orientation-Text Paired Data at Scale

Our goal is to develop an *orientation model* capable of identifying semantic orientations in open-world settings using large-scale 3D data. To support this, we introduce OrienText300K, a curated dataset of 3D models annotated with diverse language-guided orientation labels. The dataset is constructed from Objaverse [20], which contains approximately 800K Internet-sourced 3D models across a wide range of categories. Since the raw data includes noisy annotations and low-quality samples, we apply a rigorous filtering process. Using Blender, we render over 8M high-quality images under carefully designed lighting conditions to ensure fidelity for training.

Data Filtering To ensure high-quality data for generating semantic orientation annotations, we apply a dedicated filtering strategy that retains only the samples meeting the following six criteria. ❶ Standard orthogonal view only. Samples in random views will be filtered. ❷ Clean objects without the ground for auxiliary visualization. ❸ Reasonable objects that have sufficient spatial reasoning potentials. ❹ High-quality objects. Blurry and wrong samples are filtered. ❺ Distinguishable objects. Abstract and meaningless objects are filtered. ❻ Non-scene objects for object-centric understanding.

However, it is non-trivial to conduct filtering on such big data using manual labor. Inspired by recent works showing large VLMs are human-aligned judgers [147, 121, 85], we employ GPT-4o [48] by prompting requirements above. To be specific, the multi-view images of 3D objects are concatenated together with our designed prompts into GPT-4o, and GPT-4o will decide whether samples should be filtered. The filtered dataset yields 350K+ clean samples, significantly reducing data noise.

Data Annotation As mentioned in the introduction, VLMs struggle to produce accurate object orientation values, which presents a significant challenge for data generation. Fortunately, VLMs are powerful discriminators capable of distinguishing between different views through multimodal understanding. We believe that the initial stage of data cleaning effectively removed a large amount of misaligned data, leaving behind a set of properly aligned instances capable of producing *standard* orthogonal views. We then leverage GPT-4o to interpret the semantic content across six views and generate semantic-view pairs accordingly. Throughout the annotation process, both human modelers in Objaverse and ChatGPT serve as our annotators, supplying the necessary knowledge to produce both view-aligned data and semantically grounded annotations.

Quality Validation To validate annotation quality, we construct a validation set containing 208 samples with manually labeled filtering criteria and semantic orientation labels, respectively. From Fig. 3b, we observe that GPT-4o achieves an average accuracy of 88.3% and 97.1% accuracy on filtering and annotating, respectively. This provides a quality guarantee of our OrienText300K.

2.3 PointSO: A Cross-Modal 3D Transformer for Semantic Orientation Prediction

We introduce PointSO, a plain Transformer-based architecture [114] with cross-modal 3D-language fusion as our orientation model. As illustrated in Fig. 4, PointSO takes the object's 3D point clouds and a language description as inputs, and predicts the corresponding semantic orientation.

3D and Language Embeddings Given an object's point cloud $X = \{\mathbf{x}_i \in \mathbb{R}^3 | i = 1, 2, \ldots, N\}$ with N 3D points defined in (x, y, z) Cartesian space, and an arbitrary language description ℓ, we first embed both into discrete token embeddings. For the 3D point clouds, we follow [26, 136, 89] to first sample N_s seed points using farthest point sampling (FPS) and then group inputs with KNN for point feature embedding with a local geometric extraction network such as lightweight PointNet [86, 87]. An MLP head is used which maps a special `[CLS]` token [28] to a predicted direction. As for the language inputs, we adopt CLIP [97] and use the global token as cross-modal fusion inputs.

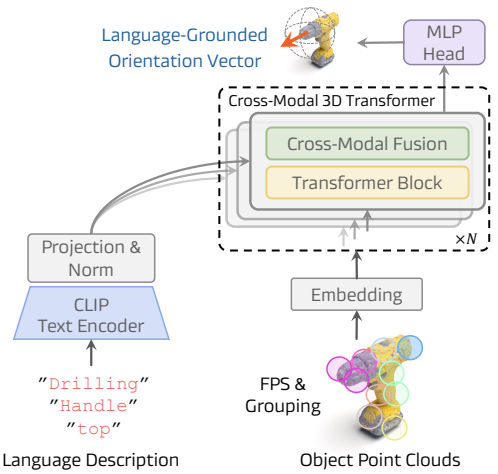


Figure 4: **PointSO model architecture**.

Cross-Modal Fusion We perform cross-modal fusion by injecting global text features into each layer of the 3D Transformer using a simple yet effective strategy: adding the text token to every point token. While other fusion methods such as cross-attention, adapters, or concatenation along spatial or channel dimensions are possible, we empirically find that token-wise addition performs best (see Appendix C.3). This effectiveness may stem from the short language inputs, where summation helps reinforce their influence across layers.

Optimization Let $\mathcal{F}_{\text{SO}}$ represent the PointSO model parameterized by θ_{SO} (the CLIP is kept frozen and thus its parameters are not included). Given every object point cloud $X_i \in \mathcal{D}_{\text{OrienText300K}}$ in the OrienText300K dataset, where each object is labeled with a language set $L_i = \{\ell_j^i, j = 1, 2, \ldots, Q\}$ and the corresponding ground truth semantic orientation set, $S_i = \{\mathbf{s}_j^i, j = 1, 2, \ldots, Q\}$. The optimization is to minimize the negative cosine similarity $\mathcal{L}_{\cos}(\mathbf{v}, \mathbf{k}) = \mathbf{1} - \frac{\mathbf{v} \cdot \mathbf{k}}{\|\mathbf{v}\| \cdot \|\mathbf{k}\|}$ between predicted and the ground truth semantic orientations:

$$\min_{\theta_{\text{SO}}} \sum_{X_i \in \mathcal{D}_{\text{OrienText300K}}} \sum_{\ell_j^i \in L_i} \mathcal{L}_{\cos}\Big(\mathcal{F}_{\text{SO}}(X_i, \ell_j^i), \mathbf{s}_j^i\Big). \tag{2}$$

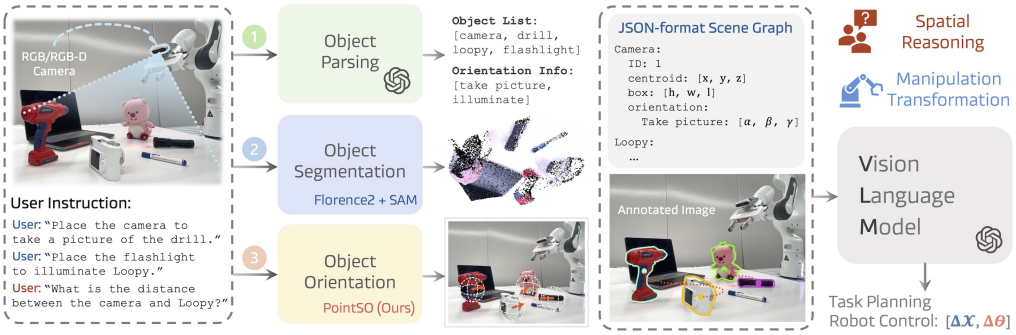


Figure 5: **Overview of SoFar system.** Given RGB-D images and language instructions, SoFar first leverages a VLM to identify relevant object phrases and semantic orientations. Then utilizes foundation models Florence-2 [125], SAM [57], and our PointSO for object segmentation and semantic orientation estimation. This information forms a 6-DoF scene graph, which the VLM uses alongside the RGB image to perform spatial understanding tasks or generate manipulation actions.

3 SoFar: Semantic Orientation Bridges Spatial Reasoning and Object Manipulation

Our proposed PointSO model now paves the way for off-the-shelf object-centric spatial orientation understanding. However, it remains challenging to extend such object-centric spatial understanding for scene-level spatial reasoning both in the digital world (*e.g.*, 6-DoF visual question answering) and in the physical world (*e.g.*, robot manipulations). To bridge this gap, we build an integrated reasoning system where a powerful VLM acts as an agent and reasons about the scene while communicating with off-the-shelf models including PointSO and SAM [57]. Fig. 5 illustrates an overview of our proposed framework, aiming at **S**emantic **O**rientation **F**or **A**utonomous **R**obots (**SoFar**).

3.1 Scene Graph with 6-DoF Information

To integrate both the positional & orientational interaction relationships of objects, we use a scene graph with 6-DoF information to represent the environment.

Position & Orientation Information Extraction Given a language query $\mathcal{Q}$, we first prompt a vision-language model $\mathcal{F}_{\text{VLM}}$ to extract a task-relevant set of object phrases $\mathcal{P} = \{p_i \,|\, i = 1, 2, \ldots, M\}$. Each phrase p_i represents a language description of an object relevant to $\mathcal{Q}$. Using the SAM [57] & Florence-2 [125], we perform language-conditioned segmentation to obtain a corresponding object set $\mathcal{X} = \{X_i \,|\, i = 1, 2, \ldots, M\}$, where X_i is the 3D point cloud of the i-th object. Each object is assigned a unique ID for use in Set-of-Mark (SoM) prompting [129]. We then prompt the VLM to generate a set of task-specific orientation descriptions L_i for related objects, and use pretrained PointSO to infer their semantic orientations, resulting in a semantic orientation set S_i.

6-DoF Scene Graph From the segmented object set $\mathcal{X}$, we construct an 6-DoF scene graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ with M nodes. Each node $\mathbf{o}_i \in \mathbf{V}$ encodes the following semantic and spatial attributes: ❶ object phrase p_i with a unique instance ID; ❷ 3D position $\mathbf{c}_i = (x, y, z) \in \mathbb{R}^3$ from the object's centroid; ❸ bounding box size $\mathbf{b}_i = (h, w, l) \in \mathbb{R}^3$; ❹ semantic orientation set S_i along with its corresponding description set L_i. Each edge $\mathbf{e}_{ij} \in \mathbf{E}$ represents the relative translation and size ratio between two connected objects $\mathbf{o}_i$ and $\mathbf{o}_j$.

3.2 Spatial-Aware Task Reasoning

We encode the 6-DoF scene graph $\mathcal{G}$ into descriptive language and input it to the VLM alongside the RGB image I and query $\mathcal{Q}$. This enriched spatial representation enables the VLM to perform accurate spatial reasoning by leveraging its visual and linguistic understanding.

Chain-of-Thought Spatial Reasoning Most robot manipulation tasks involving rigid objects can be abstracted as applying transformations to adjust their position and orientation. To guide the VLM in generating such transformations from language instructions, we adopt a CoT reasoning process [119] that decomposes the reasoning into three steps: (i) analyzing the scene with the query $\mathcal{Q}$ and object

Figure 6: **Qualitative results** of real world language-grounded manipulation. SoFar can generalize across various **embodiments, tasks and environments**.

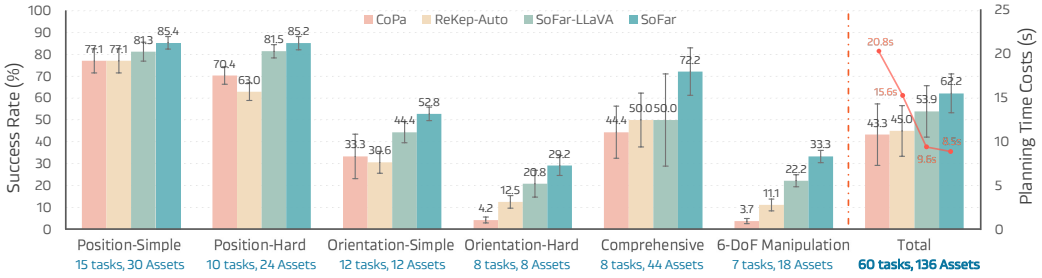


Figure 7: **Quantitative evaluation** of zero-shot real-world language-grounded rearrangement. We design **60** diverse real-world tasks involving over **100** diverse objects (detailed in Table 13).

nodes $\mathbf{V}$; (ii) computing the desired position and orientation of the target object; (iii) predicting the target position $\tilde{\mathbf{c}}_i$ and semantic orientation set $\tilde{S}_i$ for each object. Given the initial state $\mathbf{c}_i$ and S_i, the full 6-DoF transformation $\mathbf{P}_i$ is computed. Specifically, translation is obtained by $\mathbf{t}_i = \tilde{\mathbf{c}}_i - \mathbf{c}_i$, and rotation $\mathbf{R}_i$ is estimated from S_i and $\tilde{S}_i$ using the Kabsch-Umeyama algorithm [52, 53, 112].

Low-Level Motion Execution Following CoPa [44], we integrate task-specific grasping and motion planning. Object or part segmentation is performed using Florence-2 [125] and SAM [57], followed by grasp candidate generation via GSNet [33]. The optimal grasp is selected by considering both grasp quality and heuristics. Based on instruction, SoFar predicts the object's translation and rotation, defining the transformation from grasp to placement. We employ OMPL [103] to generate a collision-free trajectory, initializing joint positions at the midpoint to ensure smooth and safe motion.

Table 1: **6-DoF object rearrangement evaluation** on Open6DOR [25].

Method	Position Track			Rotation Track				6-DoF Track			Time Cost (s)
	Level 0	Level 1	Overall	Level 0	Level 1	Level 2	Overall	Position	Rotation	Overall	
Perception Tasks on Issac Sim [80] (Open6DOR V1 Setting)											
GPT-4V [81]	46.8	39.1	45.2	9.1	6.9	11.7	9.2	-	-	-	-
Dream2Real [54]	17.2	11.0	15.9	37.3	27.6	26.2	31.3	26.2	18.7	13.5	358.3s
VoxPoser [46]	35.6	21.7	32.6	-	-	-	-	-	-	-	-
Open6DOR-GPT [25]	78.6	60.3	74.9	45.7	32.5	49.8	41.1	84.8	40.0	35.6	126.3 s
SoFar-LLaVA	86.3	57.9	78.7	62.5	30.2	67.1	48.6	83.0	48.2	40.3	9.6s
SoFar	**96.0**	**81.5**	**93.0**	**68.6**	**42.2**	**70.1**	**57.0**	**92.7**	**52.7**	**48.7**	**8.5s**
Execution Tasks on Libero [64] (Open6DOR V2 Setting)											
Octo [107]	51.2	32.1	47.2	10.7	18.3	29.9	17.2	45.6	8.0	8.0	-
OpenVLA [56]	51.6	32.4	47.6	11.0	18.5	30.6	17.6	46.2	8.2	8.2	-
SoFar	**72.1**	**47.6**	**67.0**	**28.3**	**18.3**	**34.7**	**25.7**	**63.7**	**25.6**	**18.4**	**40s**

4 Experiments

4.1 Real-world Language-Grounded Object Manipulation

Tasks and Evaluations We construct 60 real-world tasks involving over 100 objects, following the Open6DOR benchmark [25]. The tasks are divided into three tracks—position, orientation, and comprehensive & 6-DoF—each with simple and hard variants. The position track assesses spatial reasoning from basic (*e.g.*, front/back/left/right) to more complex relations (*e.g.*, between/center/custom). The orientation track includes part-level orientation in the simple setting, and fine-grained angle estimation in the hard setting. The comprehensive and 6-DoF tracks evaluate complex instruction understanding and simultaneous control over position and orientation. Each task is repeated three times to ensure statistical robustness. More details and visualizations are available in Appendix D.1.

Results As shown in Fig. 7, SoFar consistently outperforms baselines across all tracks, especially on orientation and 6-DoF tasks, while maintaining low planning overhead. We also demonstrate SoFar's embodiment generality with different end-effectors, including dexterous hands and suction cups, as illustrated in Fig. 6. Additional robot setups and generalization results are provided in Appendix A.

4.2 Semantic Orientation Prediction

Using free-text descriptions to extract semantic orientations from object point clouds is challenging. In Objaverse [20], we manually annotate 128 diverse objects and construct the OrienText300K val split to evaluate the directional prediction accuracy of PointSO. We train different model variants on OrienText300K, and the results in Table 2 report performance across different angular thresholds ranging from 45° to 5°. PointSO still has an accuracy rate of 60% even under a 5° threshold.

In the real world, obtaining complete object point clouds is often difficult. To evaluate the robustness of PointSO under such conditions, we introduce three types of input perturbations: random rotations, partial single-sided observations, and Gaussian noise. As reported in Table 3, the accuracy at the 45° threshold reflects the model's resilience to these corruptions.

Table 2: **Semantic Orientation evaluation** on OrienText300K validation split.

Method	45°	30°	15°	5°	Avg.
PointSO-S	77.34	74.22	67.97	60.94	70.12
PointSO-B	**79.69**	**77.34**	**70.31**	**62.50**	**72.46**
PointSO-L	**81.25**	**78.13**	**72.66**	**65.63**	**74.42**

Table 3: **Semantic Orientation evaluation** of *robustness*. `Single-View`: randomly select a camera viewpoint within the unit sphere and generate a single FoV viewpoint in polar coordinates. `Jitter`: Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $\sigma = 0.01$. `Rotate`: random SO(3) rotation $(\alpha, \beta, \gamma) \sim \mathcal{U}(-\pi, \pi)$. `All`: all corruptions.

Method	OrienText300K-C Variants			
	`Single-View`	`Jitter`	`Rotate`	`All`
PointSO-S	72.66	76.56	73.43	67.19
PointSO-B	**75.00**	**78.90**	**75.78**	**71.09**
PointSO-L	**76.56**	**81.25**	**77.34**	**74.22**

4.3 6-DoF Object Rearrangement Evaluation on Open6DOR V2

To evaluate 6-DoF object rearrangement capabilities, we extend the original Open6DOR benchmark [25], which primarily focuses on final pose estimation, into a more comprehensive setting that

Table 4: **SimplerEnv [62] simulation evaluation results for the Google Robot setup.** We present success rates for the "Variant Aggregation" and "Visual Matching" approaches. Top-1 & Top-2 accuracies are represented using different colors. OXE: Open X-Embodiment dataset [15].

Google Robot Evaluation Setup	Policy	Training Data	Pick Coke Can				Move Near	Open / Close Drawer			Average
			Horizontal Laying	Vertical Laying	Standing	Average	Average	Open	Close	Average	
Variant Aggregation	RT-1-X [15]	OXE	0.569	0.204	0.698	0.490	0.323	0.069	**0.519**	0.294	0.397
	RT-2-X [152]	OXE	0.822	0.754	0.893	0.823	**0.792**	**0.333**	0.372	**0.353**	0.661
	Octo-Base [107]	OXE	0.005	0.000	0.013	0.006	0.031	0.000	0.021	0.011	0.012
	OpenVLA [56]	OXE	0.711	0.271	0.653	0.545	0.477	0.158	0.195	0.177	0.411
	SoFar	**Zero-Shot**	**0.861**	**0.960**	**0.901**	**0.907**	0.740	0.200	0.394	0.297	**0.676**
Visual Matching	RT-1-X [15]	OXE	**0.820**	0.330	0.550	0.567	0.317	**0.296**	**0.891**	**0.597**	0.534
	RT-2-X [152]	OXE	0.740	0.740	0.880	0.787	0.779	0.157	0.343	0.250	0.606
	Octo-Base [107]	OXE	0.210	0.210	0.090	0.170	0.042	0.009	0.444	0.227	0.168
	OpenVLA [56]	OXE	0.270	0.030	0.190	0.163	0.462	0.194	0.518	0.356	0.277
	SoFar	**Zero-Shot**	0.770	**1.000**	**1.000**	**0.923**	**0.917**	0.227	0.578	0.403	**0.749**

Table 5: **SimplerEnv [62] simulation evaluation results for the WidowX + Bridge setup.** We report both the final success rate ("Success") along with partial success (*e.g.*, "Grasp Spoon"). OXE: Open X-Embodiment dataset [15]. Bridge: BridgeData V2 dataset [115] (In domain training).

Policy	Training Data	Put Spoon on Towel		Put Carrot on Plate		Stack Green Block on Yellow Block		Put Eggplant in Yellow Basket		Average
		Grasp Spoon	Success	Grasp Carrot	Success	Grasp Green Block	Success	Grasp Eggplant	Success	
RT-1-X [6]	OXE	0.167	0.000	0.208	0.042	0.083	0.000	0.000	0.000	0.011
Octo-Base [107]	OXE	0.347	0.125	0.528	0.083	0.319	0.000	0.667	0.431	0.160
Octo-Small [107]	OXE	**0.778**	0.472	0.278	0.097	0.403	0.042	0.875	0.569	0.300
OpenVLA [56]	OXE	0.041	0.000	0.333	0.000	0.125	0.000	0.083	0.041	0.010
RoboVLM [61]	OXE	0.375	0.208	0.333	0.250	0.083	0.083	0.000	0.000	0.135
RoboVLM [61]	Bridge	0.542	0.292	0.250	0.250	0.458	0.125	0.583	0.583	0.313
SpatialVLA [94]	OXE	0.250	0.208	0.417	0.208	0.583	0.250	0.792	0.708	0.344
SpatialVLA [94]	Bridge	0.208	0.167	0.292	0.250	0.625	0.292	**1.000**	**1.000**	0.427
SoFar	**Zero-Shot**	0.625	**0.583**	**0.750**	**0.667**	**0.917**	**0.708**	0.667	0.375	**0.583**

includes both perception and execution evaluation. We migrate its scenes into a robosuite-based simulation environment [151], following the task interface defined by LIBERO [64], and name this new benchmark Open6DOR V2. Results are reported in Table 1. For perception tasks, we adopt the original Open6DOR [25] evaluation protocol and compare with the same baselines. SoFar achieves the best performance, demonstrating strong spatial understanding and zero-shot generalization. For execution tasks, we compare against the pretrained Octo [107] and the LIBERO-finetuned OpenVLA [56], all evaluated in the same robosuite environment to minimize domain shift. While both baselines show limited success due to poor generalizability, SoFar reaches around 40% success rate using a vanilla execution pipeline. We note that certain objects are intrinsically difficult to manipulate, suggesting the need for more robust policies incorporating prehensile grasping and adaptive strategies to improve performance on Open6DOR V2.

4.4 Simulation Object Manipulation Evaluation on SIMPLER [62]

We conduct quantitative evaluations of SoFar's zero-shot execution performance on Google Robot tasks & Widow-X tasks and compare it to baselines including Octo [107], OpenVLA [56] and more concurrent works [61, 94]. The robot follows the planned trajectory generated by the planning module, as described in Sec. 3.2, to execute the task. Furthermore, leveraging the error detection and re-planning capabilities of VLMs [48, 1], we can make multiple attempts following a single-step execution failure to approximately achieve a closed-loop effect. For fairness, we limit the maximum number of attempts to three. Detailed visualizations and analyses are provided in the Appendix B.5. As shown in Tables 4 and 5, despite the training data for Octo and OpenVLA including Google Robot tasks, SoFar demonstrates superior zero-shot performance compared to most baselines.

Figure 8: **Real-world orientation-aware navigation.** We present both the third-person view and the egocentric view, annotating the predicted orientation of the interacted objects.

4.5 Orientation-Aware Robotic Navigation

In navigation tasks, reaching an object from its functional side is crucial for subsequent manipulation—for example, approaching a microwave from the front to open its door. To support such scenarios, we extend *semantic orientation* to the navigation domain. As shown in Fig. 8, a quadruped robot is tasked with reaching both the correct position and the appropriate facing direction. This orientation-aware constraint enhances the navigation process by ensuring precise alignment with the desired orientation, thereby improving task performance in scenarios where directionality is critical.

4.6 Spatial Reasoning Evaluation on 6-DoF SpatialBench

To assess spatial understanding with full 6-DoF awareness, we introduce **6-DoF SpatialBench**, a VQA benchmark designed to evaluate both positional and orientational comprehension. Unlike prior benchmarks [12, 8, 29, 106] that primarily emphasize coarse positional reasoning (*e.g.*, "to the left," "nearest") and often overlook orientation or rely on relative metrics, we provide a more fine-grained evaluation with quantitative annotations. It consists of 223 human-annotated samples, each containing an RGB image and a multiple-choice question with 4 options. The benchmark includes two tracks: position and orientation, covering tasks such as object counting, spatial relations, and object-facing direction. All questions and ground-truth answers are curated through **human annotation**. We evaluate SoFar on 6-DoF SpatialBench against several VLMs and comparable methods as baselines, as presented in Table 6. SoFar consistently outperforms other methods across both tracks, achieving over 18% improvement.

Table 6: **Spatial reasoning evaluation** on 6-DoF SpatialBench. *Depth-Esti*: Use depth estimation methods such as Metric3D [135] or Moge [117].

Method	Depth-Esti	Position		Orientation		Total
		rel.	abs.	rel.	abs.	
Blind Evaluation with LLMs						
GPT-3.5-Turbo [7]	✗	24.5	24.9	26.7	27.5	25.7
GPT-4-Turbo [82]	✗	27.2	27.3	29.2	27.9	27.8
General VLMs						
LLaVA-1.5 [68]	✗	30.9	24.5	28.3	25.8	27.2
GPT-4o-mini [48]	✗	33.3	26.9	32.5	23.8	31.0
GPT-4o [48]	✗	49.4	28.4	44.2	25.8	36.2
VLMs with Spatial Awareness						
SpaceLLaVA [10]	✗	32.4	30.5	30.9	24.9	28.2
SpaceMantis [10]	✗	33.6	29.2	27.2	25.0	28.9
SpatialBot [8]	✓	50.9	21.6	39.6	22.9	32.7
RoboPoint [137]	✗	43.8	30.8	33.8	25.8	33.5
SoFar	✓	**59.6**	**33.8**	**54.6**	**31.3**	**43.9**

5 Limitations & Conclusions

One notable limitation for decoupled systems like SoFar is that the execution may fail due to a sub-module error, as shown in Appendix B.8, *i.e.*, robots may place target objects with an error transformation because of unstable grasping or inaccurate visual perception. For example, the pen will be placed in an unexpected pose due to the rotation during execution. Future works include integrating scalable data and more advanced models and exploring the potential of combining end-to-end and such decoupled methods, and expanding SoFar to more applications.

We propose *semantic orientation*, a language-grounded representation that links object orientations with intuitive descriptors (*e.g.*, "plug-in direction"), bridging geometric reasoning and functional semantics. To support this, we construct OrienText300K, a large-scale dataset of 3D models with semantic orientation annotations. Our PointSO model, integrated within the SoFar system, demonstrates strong performance in both simulated and real-world robotic manipulation tasks.

References

[1] Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805, 2023. 8, 30, 31, 33

[2] Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis K. Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 1534–1543. IEEE Computer Society, 2016. 36

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966, 2023. 29, 33

[4] Suneel Belkhale, Tianli Ding, Ted Xiao, Pierre Sermanet, Quon Vuong, Jonathan Tompson, Yevgen Chebotar, Debidatta Dwibedi, and Dorsa Sadigh. RT-H: action hierarchies using language. *CoRR*, abs/2403.01823, 2024. 36

[5] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. 32

[6] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S. Ryoo, Grecia Salazar, Pannag R. Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong T. Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: robotics transformer for real-world control at scale. In *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, 2023. 8, 36

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2020. 9, 36, 38

[8] Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *CoRR*, abs/2406.13642, 2024. 2, 9, 36

[9] Matthew Chang, Théophile Gervet, Mukul Khanna, Sriram Yenamandra, Dhruv Shah, So Yeon Min, Kavit Shah, Chris Paxton, Saurabh Gupta, Dhruv Batra, Roozbeh Mottaghi, Jitendra Malik, and Devendra Singh Chaplot. GOAT: GO to any thing. In Dana Kulic, Gentiane Venture, Kostas E. Bekris, and Enrique Coronado (eds.), *Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024*, 2024. 38

[10] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas J. Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14455–14465. IEEE, 2024. 2, 9, 36

[11] Zixuan Chen, Xialin He, Yen-Jen Wang, Qiayuan Liao, Yanjie Ze, Zhongyu Li, S. Shankar Sastry, Jiajun Wu, Koushil Sreenath, Saurabh Gupta, and Xue Bin Peng. Learning smooth humanoid locomotion through lipschitz-constrained policies. *CoRR*, abs/2410.11825, 2024. 38

[12] 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 Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 2, 9, 36

[13] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 16901–16911. IEEE, 2024. 32

[14] Xuxin Cheng, Yandong Ji, Junming Chen, Ruihan Yang, Ge Yang, and Xiaolong Wang. Expressive whole-body control for humanoid robots. In Dana Kulic, Gentiane Venture, Kostas E. Bekris, and Enrique Coronado (eds.), *Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024*, 2024. 38

[15] Open X.-Embodiment Collaboration, Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alexander Herzog, Alex Irpan, Alexander Khazatsky, Anant Raj, Anikait Singh, Anthony Brohan, Antonin Raffin, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Brian Ichter, Cewu Lu, Charles Xu, Chelsea Finn, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Chuer Pan, Chuyuan Fu, Coline Devin, Danny Driess, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Federico Ceola, Fei Xia, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Giulio Schiavi, Gregory Kahn, Hao Su, Haoshu Fang, Haochen Shi, Heni Ben Amor, Henrik I. Christensen, Hiroki Furuta, Homer Walke, Hongjie Fang, Igor Mordatch, Ilija Radosavovic, and et al. Open x-embodiment: Robotic learning datasets and RT-X models. *CoRR*, abs/2310.08864, 2023. 8, 36

[16] Jaime Corsetti, Davide Boscaini, Changjae Oh, Andrea Cavallaro, and Fabio Poiesi. Open-vocabulary object 6d pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 18071–18080. IEEE, 2024. 2

[17] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 2432–2443. IEEE Computer Society, 2017. 36

[18] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023. 29

[19] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 32, 38

[20] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 13142–13153. IEEE, 2023. 2, 3, 7, 34, 35

[21] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel R-CNN: towards high performance voxel-based 3d object detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 1201–1209. AAAI Press, 2021. 36

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. 36

[23] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. PLA: language-driven open-vocabulary 3d scene understanding. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 37

[24] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Lowis3d: Language-driven open-world instance-level 3d scene understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46 (12):8517–8533, 2024. 37

[25] Yufei Ding, Haoran Geng, Chaoyi Xu, Xiaomeng Fang, Jiazhao Zhang, Songlin Wei, Qiyu Dai, Zhizheng Zhang, and He Wang. Open6dor: Benchmarking open-instruction 6-dof object rearrangement and A vlm-based approach. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2024, Abu Dhabi, United Arab Emirates, October 14-18, 2024*, pp. 7359–7366. IEEE, 2024. 7, 8, 32, 33, 34, 36, 37

[26] Runpei Dong, Zekun Qi, Linfeng Zhang, Junbo Zhang, Jianjian Sun, Zheng Ge, Li Yi, and Kaisheng Ma. Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? In *Int. Conf. Learn. Represent. (ICLR)*, 2023. 2, 4, 33, 36, 37, 38

[27] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. DreamLLM: Synergistic multimodal comprehension and creation. In *Int. Conf. Learn. Represent. (ICLR)*, 2024. 36

[28] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Int. Conf. Learn. Represent. (ICLR)*, 2021. 4, 33

[29] 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, ACL 2024 - Short Papers, Bangkok, Thailand, August 11-16, 2024*, pp. 346–355. Association for Computational Linguistics, 2024. 9, 29

[30] Ben Eisner, Harry Zhang, and David Held. Flowbot3d: Learning 3d articulation flow to manipulate articulated objects. In Kris Hauser, Dylan A. Shell, and Shoudong Huang (eds.), *Robotics: Science and Systems XVIII, New York City, NY, USA, June 27 - July 1, 2022*, 2022. 29

[31] Nico Engel, Vasileios Belagiannis, and Klaus Dietmayer. Point transformer. *IEEE Access*, 9:134826–134840, 2021. 36

[32] Guofan Fan, Zekun Qi, Wenkai Shi, and Kaisheng Ma. Point-gcc: Universal self-supervised 3d scene pre-training via geometry-color contrast. In Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (eds.), *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pp. 4709–4718. ACM, 2024. 37

[33] Haoshu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 11441–11450. Computer Vision Foundation / IEEE, 2020. 6

[34] Kuan Fang, Fangchen Liu, Pieter Abbeel, and Sergey Levine. MOKA: Open-World Robotic Manipulation through Mark-Based Visual Prompting. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. 36

[35] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *CoRR*, abs/2403.11401, 2024. 36

[36] James J Gibson. The theory of affordances:(1979). In *The people, place, and space reader*, pp. 56–60. Routledge, 2014. 37

[37] Mohit Goyal, Sahil Modi, Rishabh Goyal, and Saurabh Gupta. Human hands as probes for interactive object understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 3283–3293. IEEE, 2022. 37

[38] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. MVTN: multi-view transformation network for 3d shape recognition. In *Int. Conf. Comput. Vis. (ICCV)*, pp. 1–11. IEEE, 2021. 36

[39] Jiawei He, Danshi Li, Xinqiang Yu, Zekun Qi, Wenyao Zhang, Jiayi Chen, Zhaoxiang Zhang, Zhizheng Zhang, Li Yi, and He Wang. Dexvlg: Dexterous vision-language-grasp model at scale. *CoRR*, abs/2507.02747, 2025. 36

[40] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022. 36, 38

[41] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *Annu. Conf. Robot. Learn. (CoRL)*, 2024. 38

[42] Xialin He, Runpei Dong, Zixuan Chen, and Saurabh Gupta. Learning getting-up policies for real-world humanoid robots. *CoRR*, abs/2502.12152, 2025. 38

[43] 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 Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 36

[44] Haoxu Huang, Fanqi Lin, Yingdong Hu, Shengjie Wang, and Yang Gao. Copa: General robotic manipulation through spatial constraints of parts with foundation models. *CoRR*, abs/2403.08248, 2024. 6, 28, 34, 36

[45] Siyuan Huang, Haonan Chang, Yuhan Liu, Yimeng Zhu, Hao Dong, Abdeslam Boularias, Peng Gao, and Hongsheng Li. A3VLM: actionable articulation-aware vision language model. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, volume 270 of *Proceedings of Machine Learning Research*, pp. 1675–1690. PMLR, 2024. 36

[46] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. In *Annu. Conf. Robot. Learn. (CoRL)*, 2023. 7, 36, 39

[47] Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. In *Annu. Conf. Robot. Learn. (CoRL)*, 2024. 28, 30, 34, 37, 38, 39

[48] Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll L. Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, and Dane Sherburn. Gpt-4o system card. *CoRR*, abs/2410.21276, 2024. 2, 4, 8, 9, 30, 31

[49] Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J. Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. Do as I can, not as I say: Grounding language in robotic affordances. In *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, volume 205 of *Proceedings of Machine Learning Research*, pp. 287–318. PMLR, 2022. 2, 36, 37

[50] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, and Ilge Akkaya. Openai o1 system card. *CoRR*, abs/2412.16720, 2024. 31

[51] Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *CoRR*, abs/2506.03135, 2025. 36

[52] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976. 6

[53] Wolfgang Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 34 (5):827–828, 1978. 6

[54] Ivan Kapelyukh, Yifei Ren, Ignacio Alzugaray, and Edward Johns. Dream2real: Zero-shot 3d object rearrangement with vision-language models. In *IEEE International Conference on Robotics and Automation, ICRA 2024, Yokohama, Japan, May 13-17, 2024*, pp. 4796–4803. IEEE, 2024. 7

[55] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R. Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, and et al. DROID: A large-scale in-the-wild robot manipulation dataset. *CoRR*, abs/2403.12945, 2024. 36

[56] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Paul Foster, Grace Lam, Pannag 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. *CoRR*, abs/2406.09246, 2024. 7, 8, 36

[57] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 3992–4003. IEEE, 2023. 2, 5, 6, 30, 32

[58] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. In *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, volume 205 of *Proceedings of Machine Learning Research*, pp. 715–725. PMLR, 2022. 2

[59] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. *CoRR*, abs/2301.12597, 2023. 29

[60] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 18061–18070. IEEE, 2024. 2, 29, 36

[61] Xinghang Li, Peiyan Li, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Tao Kong, Hanbo Zhang, and Huaping Liu. Towards generalist robot policies: What matters in building vision-language-action models. *CoRR*, abs/2412.14058, 2024. 8

[62] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Oier Mees, Karl Pertsch, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, volume 270 of *Proceedings of Machine Learning Research*, pp. 3705–3728. PMLR, 2024. 8, 28, 30, 37

[63] 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 2023, London, UK, May 29 - June 2, 2023*, pp. 9493–9500. IEEE, 2023. 36

[64] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: benchmarking knowledge transfer for lifelong robot learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 7, 8, 28, 38

[65] Dingning Liu, Xiaomeng Dong, Renrui Zhang, Xu Luo, Peng Gao, Xiaoshui Huang, Yongshun Gong, and Zhihui Wang. 3daxiesprompts: Unleashing the 3d spatial task capabilities of GPT-4V. *CoRR*, abs/2312.09738, 2023. 36

[66] Hao Liu, Lisa Lee, Kimin Lee, and Pieter Abbeel. Instruction-following agents with jointly pre-trained vision-language models. *CoRR*, abs/2210.13431, 2022. 36

[67] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023. 29, 34, 36

[68] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2024. 9, 36

[69] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVII*, volume 15105 of *Lecture Notes in Computer Science*, pp. 38–55. Springer, 2024. 31, 32

[70] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. HOI4D: A 4d egocentric dataset for category-level human-object interaction. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022. 38

[71] Yunze Liu, Changxi Chen, and Li Yi. Interactive humanoid: Online full-body motion reaction synthesis with social affordance canonicalization and forecasting. *CoRR*, abs/2312.08983, 2023. 38

[72] Yunze Liu, Junyu Chen, Zekai Zhang, Jingwei Huang, and Li Yi. Leaf: Learning frames for 4d point cloud sequence understanding. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 604–613. IEEE, 2023. 38

[73] Yunze Liu, Changxi Chen, Zifan Wang, and Li Yi. Crossvideo: Self-supervised cross-modal contrastive learning for point cloud video understanding. In *IEEE International Conference on Robotics and Automation, ICRA 2024, Yokohama, Japan, May 13-17, 2024*, pp. 12436–12442. IEEE, 2024. 38

[74] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 2929–2938. IEEE, 2021. 36

[75] Chenyang Ma, Kai Lu, Ta Ying Cheng, Niki Trigoni, and Andrew Markham. Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3d priors. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 36

[76] Jiageng Mao, Yujing Xue, Minzhe Niu, Haoyue Bai, Jiashi Feng, Xiaodan Liang, Hang Xu, and Chunjing Xu. Voxel transformer for 3d object detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 3144–3153. IEEE, 2021. 36

[77] Weixin Mao, Weiheng Zhong, Zhou Jiang, Dong Fang, Zhongyue Zhang, Zihan Lan, Fan Jia, Tiancai Wang, Haoqiang Fan, and Osamu Yoshie. Robomatrix: A skill-centric hierarchical framework for scalable robot task planning and execution in open-world. *CoRR*, abs/2412.00171, 2024. 36

[78] Daniel Maturana and Sebastian A. Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *IEEE/RSJ Int. Conf. Intell. Robot. and Syst. (IROS)*, pp. 922–928. IEEE, 2015. 36

[79] Kaichun Mo, Leonidas J. Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 6793–6803. IEEE, 2021. 29

[80] NVDIA. Nvidia isaac sim. https://developer.nvidia.com/isaac-sim, 2021. 7

[81] OpenAI. Gpt-4v(ision) system card, 2023. URL https://openai.com/research/gpt-4v-system-card. 7, 29

[82] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 9

[83] Yatian Pang, Wenxiao Wang, Francis E. H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *Eur. Conf. Comput. Vis. (ECCV)*, 2022. 36

[84] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas A. Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 37

[85] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *CoRR*, abs/2406.16855, 2024. 4

[86] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 77–85, 2017. 4, 33, 36

[87] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Adv. Neural Inform. Process. Syst. (NIPS)*, pp. 5099–5108, 2017. 4, 36

[88] William Qi, Ravi Teja Mullapudi, Saurabh Gupta, and Deva Ramanan. Learning to move with affordance maps. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 37

[89] Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *Int. Conf. Mach. Learn. (ICML)*, 2023. 2, 4, 33, 36, 38

[90] Zekun Qi, Muzhou Yu, Runpei Dong, and Kaisheng Ma. VPP: efficient conditional 3d generation via voxel-point progressive representation. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023. 36

[91] Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLIII*, volume 15101 of *Lecture Notes in Computer Science*, pp. 214–238. Springer, 2024. 2, 33, 36

[92] Zhangyang Qi, Ye Fang, Zeyi Sun, Xiaoyang Wu, Tong Wu, Jiaqi Wang, Dahua Lin, and Hengshuang Zhao. Gpt4point: A unified framework for point-language understanding and generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 26407–26417. IEEE, 2024. 36

[93] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Abed Al Kader Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022. 36

[94] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, and Xuelong Li. Spatialvla: Exploring spatial representations for visual-language-action model. *CoRR*, abs/2501.15830, 2025. 8

[95] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI blog*, 2018. 36

[96] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 36

[97] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn. (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. 4, 36, 37, 38

[98] Konstantinos I. Roumeliotis and Nikolaos D. Tselikas. Chatgpt and open-ai models: A preliminary review. *Future Internet*, 15(6):192, 2023. 33

[99] Kenneth Shaw, Ananye Agarwal, and Deepak Pathak. LEAP hand: Low-cost, efficient, and anthropomorphic hand for robot learning. In Kostas E. Bekris, Kris Hauser, Sylvia L. Herbert, and Jingjin Yu (eds.), *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, 2023. 29

[100] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, volume 205 of *Proceedings of Machine Learning Research*, pp. 785–799. PMLR, 2022. 36

[101] Tianmin Shu, Michael S. Ryoo, and Song-Chun Zhu. Learning social affordance for human-robot interaction. In *Int. Joint Conf. Artif. Intell. (IJCAI)*, 2016. 37

[102] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik G. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Int. Conf. Comput. Vis. (ICCV)*, 2015. 36

[103] Ioan Alexandru Sucan, Mark Moll, and Lydia E. Kavraki. The open motion planning library. *IEEE Robotics Autom. Mag.*, 19(4):72–82, 2012. 6

[104] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. Onepose: One-shot object pose estimation without CAD models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 6815–6824. IEEE, 2022. 2

[105] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 36

[106] Emilia Szymanska, Mihai Dusmanu, Jan-Willem Buurlage, Mahdi Rad, and Marc Pollefeys. Space3d-bench: Spatial 3d question answering benchmark. *CoRR*, abs/2408.16662, 2024. 9

[107] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, 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. *CoRR*, abs/2405.12213, 2024. 7, 8

[108] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth J. Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI Conf. Artif. Intell. (AAAI)*, 2011. 2

[109] Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. Robots that use language. *Annu. Rev. Control. Robotics Auton. Syst.*, 3:25–55, 2020. 37

[110] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. 34, 36

[111] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew

Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023. 34, 36

[112] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(4):376–380, 1991. 6

[113] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 1588–1597. IEEE, 2019. 36

[114] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst. (NIPS)*, pp. 5998–6008, 2017. 2, 4, 36

[115] Homer Rich Walke, Kevin Black, Tony Z. Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, Abraham Lee, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata V2: A dataset for robot learning at scale. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, volume 229 of *Proceedings of Machine Learning Research*, pp. 1723–1736. PMLR, 2023. 8

[116] Chenxi Wang, Haoshu Fang, Minghao Gou, Hongjie Fang, Jin Gao, and Cewu Lu. Graspness discovery in clutters for fast and accurate grasp detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 15944–15953. IEEE, 2021. 29

[117] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. *CoRR*, abs/2410.19115, 2024. 9

[118] Zehan Wang, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient anything: Learning robust object orientation estimation from rendering 3d models. *CoRR*, abs/2412.18605, 2024. 38

[119] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022. 2, 5, 38

[120] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 17868–17879. IEEE, 2024. 2

[121] Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas J. Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v(ision) is a human-aligned evaluator for text-to-3d generation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2024. 4

[122] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 1912–1920. IEEE Computer Society, 2015. 36

[123] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A simulated part-based interactive environment. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020. 29

[124] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 26-30, 2018*, 2018. 2

[125] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 4818–4829. IEEE, 2024. 5, 6, 30, 31, 32

[126] Saining Xie, Jiatao Gu, Demi Guo, Charles R. Qi, Leonidas J. Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Eur. Conf. Comput. Vis. (ECCV)*, volume 12348 of *Lecture Notes in Computer Science*, pp. 574–591. Springer, 2020. 36

[127] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. *CoRR*, abs/2308.16911, 2023. 36

[128] Zhenjia Xu, Zhanpeng He, and Shuran Song. Universal manipulation policy network for articulated objects. *IEEE Robotics Autom. Lett.*, 7(2):2447–2454, 2022. 29

[129] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. *CoRR*, abs/2310.11441, 2023. 5, 34, 36

[130] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *CoRR*, abs/2412.14171, 2024. 38

[131] Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Leslie Pack Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 36

[132] Sriram Yenamandra, Arun Ramachandran, Karmesh Yadav, Austin S. Wang, Mukul Khanna, Théophile Gervet, Tsung-Yen Yang, Vidhi Jain, Alexander Clegg, John M. Turner, Zsolt Kira, Manolis Savva, Angel X. Chang, Devendra Singh Chaplot, Dhruv Batra, Roozbeh Mottaghi, Yonatan Bisk, and Chris Paxton. Homerobot: Open-vocabulary mobile manipulation. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, volume 229 of *Proceedings of Machine Learning Research*, pp. 1975–2011. PMLR, 2023. 38

[133] Li Yi, Vladimir G. Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas J. Guibas. A scalable active framework for region annotation in 3d shape collections. *ACM Trans. Graph.*, 35(6):210:1–210:12, 2016. 36

[134] Li Yi, Hao Su, Xingwen Guo, and Leonidas J. Guibas. Syncspeccnn: Synchronized spectral CNN for 3d shape segmentation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017. 36

[135] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from A single image. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 9009–9019. IEEE, 2023. 9

[136] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022. 4, 36

[137] 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 for robotics. *CoRR*, abs/2406.10721, 2024. 9, 36

[138] Zhecheng Yuan, Tianming Wei, Shuiqi Cheng, Gu Zhang, Yuanpei Chen, and Huazhe Xu. Learning to manipulate anywhere: A visual generalizable framework for reinforcement learning. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, volume 270 of *Proceedings of Machine Learning Research*, pp. 1815–1833. PMLR, 2024. 36

[139] Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. *CoRR*, abs/2412.06224, 2024. 28

[140] Junbo Zhang, Runpei Dong, and Kaisheng Ma. CLIP-FO3D: learning free open-world 3d scene representations from 2d dense CLIP. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, pp. 2040–2051. IEEE, 2023. 37

[141] Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. Llama-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 36

[142] Shaochen Zhang, Zekun Qi, Runpei Dong, Xiuxiu Bai, and Xing Wei. Positional prompt tuning for efficient 3d representation learning. *CoRR*, abs/2408.11567, 2024. 37

[143] Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, Li Yi, Wenjun Zeng, and Xin Jin. Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge. *CoRR*, abs/2507.04447, 2025. 36

[144] Liang Zhao, En Yu, Zheng Ge, Jinrong Yang, Haoran Wei, Hongyu Zhou, Jianjian Sun, Yuang Peng, Runpei Dong, Chunrui Han, and Xiangyu Zhang. Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning. In *Int. Joint Conf. Artif. Intell. (IJCAI)*, 2024. 36

[145] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, 2023. 36

[146] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. 36

[147] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 4

[148] Chengliang Zhong, Yuhang Zheng, Yupeng Zheng, Hao Zhao, Li Yi, Xiaodong Mu, Ling Wang, Pengfei Li, Guyue Zhou, Chao Yang, Xinliang Zhang, and Jian Zhao. 3d implicit transporter for temporally consistent keypoint discovery. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 3846–3857. IEEE, 2023. 29

[149] Enshen Zhou, Qi Su, Cheng Chi, Zhizheng Zhang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, and He Wang. Code-as-monitor: Constraint-aware visual programming for reactive and proactive robotic failure detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6919–6929, 2025. 36

[150] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *Int. Conf. Learn. Represent. (ICLR)*, 2024. 29

[151] Yuke Zhu, Josiah Wong, Ajay Mandlekar, and Roberto Martín-Martín. robosuite: A modular simulation framework and benchmark for robot learning. *CoRR*, abs/2009.12293, 2020. 8

[152] 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 T. 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 *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 2023. 8, 36

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We state the contributions in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations at Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: N/A

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed experiment information in the Appendix D for reproduction.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the data and code in the supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide detailed hyperparameters in the Appendix D for reproduction.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: In the real-world experiment depicted in Fig. 7, we report error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide detailed computer resources in the Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The paper has ensured anonymity and ethical standards.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss potential positive societal impacts and negative societal impacts in Appendix I.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: N/A

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In Section 4, we properly credit all the public baselines and datasets utilized in this paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We include the documentation in the supplementary materials.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: N/A

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: N/A

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: N/A

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Robot Setups

A.1 Simulation Robot Setups

To ensure fairness, we utilize the same Franka Panda arm for evaluations in both the LIBERO [64] and our Open6DOR V2 benchmarks. For SIMPLER [62], we use the Google Robot and Widow-X exclusively to conduct the baseline experiments, adhering to all configurations outlined in SIMPLER, as presented in Tables 4 and 5.

A.2 Real World Robot Setups

As for manipulation tasks, in Fig. 9, we perform 6-DoF rearrangement tasks using the Franka Panda equipped with a gripper and the UR robot arm with a LeapHand, while articulated object manipulation is conducted using the Flexiv arm equipped with a suction tool. All the robot arms mount a RealSense D415 camera at their end for image capturing.

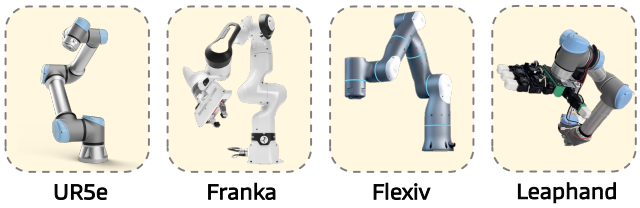


Figure 9: **The robots used in our real-world experiments.**

In Fig. 10, we present the workspace and robotic arm for real-world 6-DoF rearrangement. Unlike Rekep [47], CoPa [44] et al., we utilize only a single RealSense D415 camera. This setup significantly reduces the additional overhead associated with environmental setup and multi-camera calibration, and it is more readily reproducible.

As for navigation tasks, we provide a visualization of our robotic dog in Fig. 11. Following Uni-Navid [139], our robotic dog is Unitree GO2 and we mount a RealSense D455 camera on the head of the robotic dog. Here, we only use the RGB frames with a resolution of 640×480 in the setting of 90° HFOV. We also mount a portable Wi-Fi at the back of the robot dog, which is used to communicate with the remote server (send captured images and receive commands). Unitree GO2 is integrated with a LiDAR-L1, which is only used for local motion planning.

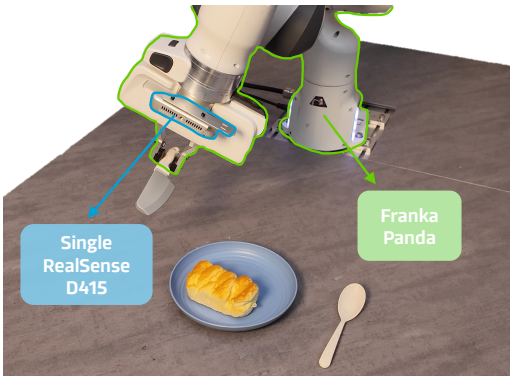


Figure 10: **6-DoF rearrangement robot setup.**

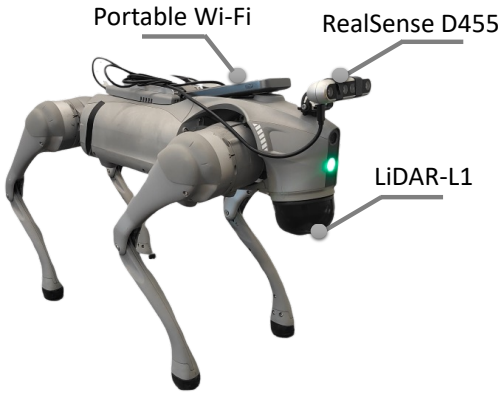


Figure 11: **Navigation robot setup.**

Table 7: **Zeroshot articulate object manipulation evaluation** within the SAPIEN [123] simulator using PartNet-Mobility Dataset. Notably, while the baseline methods use distinct training and testing splits, our model achieves robust performance without fine-tuning on the SAPIEN samples.

Method																
Where2Act [79]	0.26	0.36	0.19	0.27	0.23	0.11	0.15	0.47	0.14	0.24	0.13	0.12	0.56	0.68	0.07	0.40
UMPNet [128]	0.46	0.43	0.15	0.28	0.54	0.32	0.28	0.56	0.44	0.40	0.10	0.23	0.18	0.54	0.20	0.42
FlowBot3D [30]	0.67	0.55	0.20	0.32	0.27	0.31	0.61	**0.68**	0.15	0.28	0.36	0.18	0.21	0.70	0.18	0.26
Implicit3D [148]	0.53	0.58	0.35	0.55	0.28	**0.66**	0.58	0.51	0.52	0.57	0.45	0.34	0.41	0.54	0.39	0.43
ManipLLM [60]	0.68	0.64	0.36	0.77	0.43	0.62	0.65	0.61	0.65	0.52	0.53	**0.40**	0.64	0.71	**0.60**	**0.64**
SoFar	**0.75**	**0.88**	**0.43**	**0.85**	**0.60**	0.54	**0.75**	0.49	**0.58**	**0.72**	**0.69**	0.42	**0.70**	**0.81**	0.58	0.63

Method					AVG											AVG
Where2Act [79]	0.13	0.18	0.13	0.40	0.26	0.18	0.35	0.38	0.28	0.05	0.21	0.17	0.20	0.15	0.15	0.21
UMPNet [128]	0.22	0.33	0.26	0.64	0.35	0.42	0.20	0.35	0.42	0.29	0.20	0.26	0.28	0.25	0.15	0.28
FlowBot3D [30]	0.17	0.53	0.29	0.42	0.37	0.23	0.10	0.60	0.39	0.27	0.42	0.28	0.51	0.13	0.23	0.32
Implicit3D [148]	0.27	0.65	0.20	0.33	0.46	0.45	0.17	0.80	0.53	0.15	0.69	0.41	0.31	0.30	0.31	0.41
ManipLLM [60]	**0.41**	**0.75**	0.44	0.67	0.59	0.38	0.22	0.81	**0.86**	0.38	**0.85**	0.42	**0.83**	0.26	0.38	0.54
SoFar	0.35	0.68	**0.62**	**0.73**	**0.64**	**0.68**	**0.45**	**0.90**	0.77	**0.55**	0.79	**0.48**	0.80	**0.56**	**0.44**	**0.64**

B Additional Experiments

B.1 Articulated Objects Manipulation Evaluation

We further integrate SoFar with articulated object manipulation, as illustrated in Table 7, and evaluate its practicality in robotic manipulation tasks using the PartNet-Mobility Dataset within the SAPIEN [123] simulator. Our experimental setup follows ManipLLM [60], employing the same evaluation metrics. Specifically, we directly utilize the segmentation centers provided by SAM as contact points, leverage PointSO to generate contact directions, and use VLM to determine subsequent motion directions. The results demonstrate significant improvements over the baseline. Notably, our model achieves this performance without dividing the data into training and testing sets, operating instead in a fully zero-shot across most tasks. This underscores the robustness and generalization of our approach.

B.2 Spatial Reasoning on EmbSpatial-Bench [29]

To further validate the spatial reasoning capabilities of SoFar, we evaluated its performance on the spatial visual-question-answering tasks in EmbSpatial-Bench [29]. As reported in Table 8, our model substantially outperforms all baseline methods, achieving over a 20% improvement in overall performance. This result highlights SoFar's effectiveness in spatial understanding and reasoning within complex visual scenes.

Table 8: Evaluation of EmbSpatial-Bench [29].

Model	Generation	Likelihood
BLIP-2 [59]	37.99	35.71
InstructBLIP [18]	38.85	33.41
MiniGPT4 [150]	23.54	31.70
LLaVA-1.6 [67]	35.19	38.84
GPT-4V [81]	36.07	-
Qwen-VL-Max [3]	49.11	-
SoFar	**70.88**	-

B.3 Cross Embodiment Generalization

Our approach determines grasp poses by generating masks and plans the target pose and transformation using our PointSO and large language model. It does not rely on trajectory data specific to any robotic arm, making SoFar embodiment-agnostic. Fig. 6 illustrates the diverse embodiments employed in our real-world experiments. Leveraging the GSNet [116] algorithm based on LeapHand [99], we perform 6-DoF object manipulation experiments on dexterous hands. We conduct three position-related and three rotation-related experiments. Leveraging the PointSO and large language models, SoFar is capable of performing complex 6-DoF manipulation tasks, such as "*Upright the fallen wine glass and arrange it neatly in a row with the other wine glasses.*"

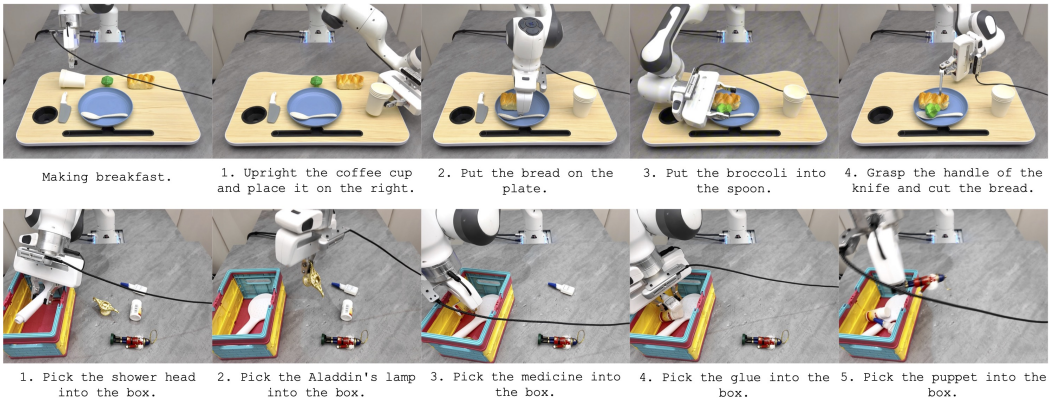


Figure 12: **Long-horizon object manipulation** experiment of our SoFar.

B.4 Long Horizon Object Manipulation Experiment

Fig. 12 illustrates the execution performance of our model on long-horizon tasks. Through the VLM [48, 1], complex instructions such as "making breakfast" and "cleaning up the desktop" can be decomposed into sub-tasks. In the second example, we deliberately chose uncommon objects as assets, such as "Aladdin's lamp" and "puppets", but SoFar is able to successfully complete all tasks.

B.5 Close-Loop Execution Experiment

Similar to ReKep [47], SoFar leverages VLMs [48, 1] to perform long-horizon decomposition of complex tasks and employs dual-system VLMs [48, 1] to determine the success of execution between tasks and subtasks, enabling closed-loop execution. When a discrepancy between the results and expectations is detected, SoFar re-percepts and re-executes the current subtask. We demonstrate the closed-loop replan capabilities of SoFar within SimplerEnv [62] in Fig. 13. The instruction for both tasks is "pick the coke can" In Fig. 13 (a), the model initially misidentified the coke can as a Fanta can. After correction by the VLM, the model re-identified and located the correct object. In Fig. 13 (b), the model accidentally knocks over the Coke can during motion due to erroneous motion planning. Subsequently, the model re-plans and successfully achieves the grasp.

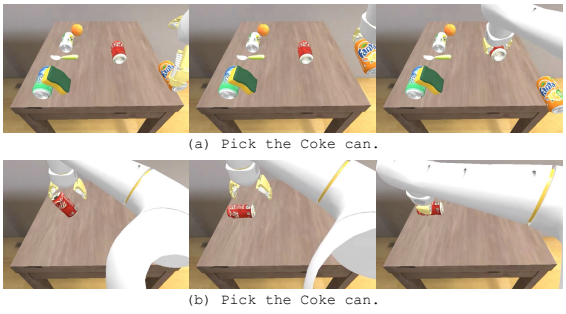


Figure 13: **Close-loop execution of our SoFar.**

B.6 In the Wild Evaluation of Semantic Orientation

We provide a qualitative demonstration of the accuracy of PointSO under in-the-wild conditions, as shown in Fig. 14, where the predicted Semantic Orientation is marked in the images. We obtained single-sided point clouds by segmenting objects using Florence-2 [125] and SAM [57] and fed them into PointSO. It can be observed that our model achieves good performance across different views, objects, and instructions, which proves the effectiveness and generalization of PointSO.

B.7 Cross-View Generalization

SoFar gets point clouds in the world coordinate system using an RGB-D camera to obtain grasping poses, and it is not limited to a fixed camera perspective. In addition, PointSO generates partial point clouds from different perspectives through random camera views to serve as data augmentation for training data, which also generalizes to camera perspectives in the real world. Fig. 15 illustrates SoFar's generalization capability for 6-DoF object manipulation across different camera poses. It can be observed that whether it's a front view, side view, or ego view, SoFar can successfully execute the "upright the bottle" instruction.

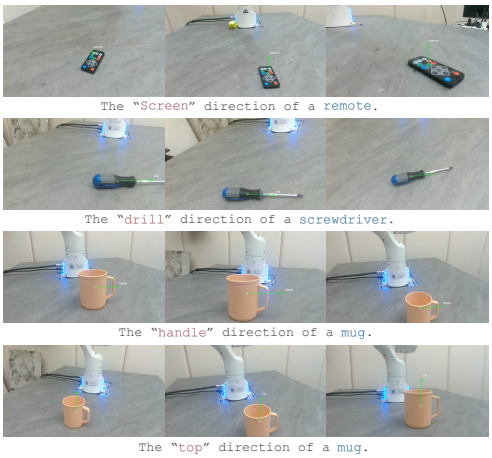


Figure 14: **In-the-wild evaluation** of PointSO.

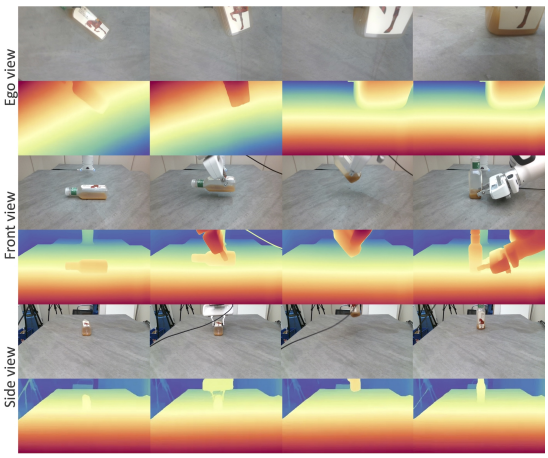


Figure 15: **Cross view generalization** of SoFar.

B.8 Failure Case Distribution Analysis

Based on the failure cases from real-world experiments, we conducted a quantitative analysis of the failure case distribution for SoFar, with the results shown in Fig. 16. It can be observed that 31% of the failures originated from grasping issues, including objects being too small, inability to generate reasonable grasping poses, and instability after grasping leading to sliding or dropping. Next, 23% were due to incorrect or inaccurate Semantic Orientation prediction. For tasks such as upright or upside - down, highly precise angle estimation (<5°) is required for smooth execution. Object analysis and detection accounted for approximately 20% of the errors. The instability of open-vocabulary detection modules like Florence2 [125] and Grounding DINO [69] often led to incorrect detection of out-of-distribution objects or object parts. In addition, since our Motion Planning did not take into account the working space range of the robotic arm and potential collisions of the manipulated object, occasional deadlocks and collisions occurred during motion. Finally, there were issues with the Task Planning of the VLM [48, 1]. For some complex Orientations, the VLM occasionally failed to infer the required angles and directions to complete the task. Employing a more powerful, thought-enabled VLM [50] might alleviate such errors.

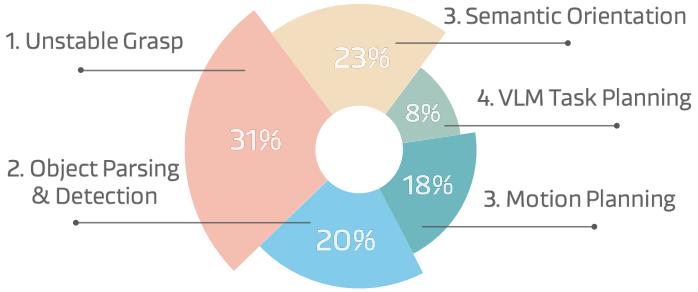


Figure 16: **Failure case distribution analysis** of our SoFar.

C Ablation Study

C.1 Semantic Orientation Ablation

To demonstrate that the proposed semantic orientation indeed plays a crucial role in robotic tasks—rather than the observed effects being attributable to other factors such as Chain-of-Thought reasoning—we conduct conduct ablation experiments for methodological differences between the baselines, including whether to add semantic orientation in the scene graph and whether to use CoT, as shown in Tab. 9.

Table 9: **Ablation study of composition module** of SoFar.

CoT	Orient.	Position Track			Rotation Track				6-DoF Track		
		Level 0	Level 1	Overall	Level 0	Level 1	Level 2	Overall	Position	Rotation	Overall
✗	✗	95.4	77.7	91.9	17.2	8.4	11.4	13.0	92.7	15.5	14.2
✓	✗	**96.3**	**81.6**	**93.3**	16.3	8.9	11.0	12.9	**93.0**	15.1	13.7
✗	✓	95.6	77.2	91.7	63.3	35.4	61.8	52.3	92.7	48.3	45.8
✓	✓	96.0	81.5	93.0	**68.6**	**42.2**	**70.1**	**57.0**	92.7	**52.7**	**48.7**

C.2 Scaling Law

The scaling capability of models and data is one of the most critical attributes today and a core feature of foundation models [5]. We investigate the performance of PointSO across different data scales, as illustrated in Table 10. We obtain the subset for OrienText300K from Objaverse-LVIS, which consists of approximately 46,000 3D objects with category annotations. The selection was based on the seven criteria mentioned in the main text. Objects meeting all seven criteria formed the strict subset, comprising around 15k objects. When including objects without textures and those of lower quality, the total increases to approximately 26k objects. It can be seen that the increase in data volume is the most significant factor driving the performance improvement of PointSO. It can be anticipated that with further data expansion, such as Objaverse-XL [19], PointSO will achieve better performance.

Table 10: **Data scaling property** of semantic orientation with different training data scales evaluated on OrienText300K validation split. All experiments are conducted with the PointSO-Base variant.

Data Scale	45°	30°	15°	5°	Average
15K	57.03	46.09	39.84	27.34	42.58
35K	61.72	53.13	43.75	30.47	47.27
150K	76.56	72.66	66.41	56.25	67.97
350K	**79.69**	**77.34**	**70.31**	**62.50**	**72.46**

C.3 Cross-Modal Fusion Choices

We further conduct an ablation study on the multi-modal fusion methods in PointSO, testing commonly used feature fusion techniques such as cross-attention, multiplication, addition, and concatenation, as shown in Table 11. The results indicate that simple addition achieves the best performance. This may be attributed to the fact that instructions in the semantic domain are typically composed of short phrases or sentences, and the text CLS token already encodes sufficiently high-level semantic information.

Table 11: **Ablation study of multi-modal fusion** in PointSO. All experiments are conducted with the PointSO-Base variant.

Fusion Method	45°	30°	15°	5°	Avg.
Cross-attn	74.22	70.31	63.28	57.03	66.21
Multiplication	74.22	69.53	60.16	56.25	65.04
Addition	**79.69**	**77.34**	**70.31**	**62.50**	**72.46**
Concat	66.41	60.94	52.34	43.75	55.86

Table 12: **Ablation study of open vocabulary detection modules** on Open6DOR perception tasks.

Method	Position Track			Rotation Track				6-DoF Track			Time Cost (s)
	Level 0	Level 1	Overall	Level 0	Level 1	Level 2	Overall	Position	Rotation	Overall	
YOLO-World [13]	59.0	37.7	53.3	48.3	36.1	62.0	44.9	53.4	44.6	27.8	**7.4s**
Grounding DINO [69]	92.2	71.5	86.7	64.7	41.1	69.8	55.5	87.2	51.6	44.6	9.2s
Florence-2 [125]	**96.0**	**81.5**	**93.0**	**68.6**	**42.2**	**70.1**	**57.0**	**92.7**	**52.7**	**48.7**	8.5s

C.4 Open Vocabulary Object Detection Module

SoFar utilize an open vocabulary detection foundation model to localize the interacted objects or parts, then generate masks with SAM [57]. Although not the SOTA performance on the COCO benchmark, Florence-2 [125] exhibits remarkable generalization in in-the-wild detection tasks, even in simulator scenarios. Table 12 illustrates the performance of various detection modules in Open6DOR [25] Perception, where Florence-2 achieves the best results and outperforms Grounding DINO [69] and YOLO-World [13].

Figure 17: **The real-world assets used in our real-world experiments**. More than 100 diverse objects are used in our 6-DoF rearrangement experiments.

D Additional Implementation Details

D.1 Detail Real World Experiment Results

To fully demonstrate the generalization of SoFar rather than cherry-picking, we carefully design 60 different real-world experimental tasks, covering more than 100 different and diverse objects. Similar to the Open6DOR [25] benchmark in the simulator, we divide these 60 tasks into three parts: position-track, orientation-track, and the most challenging comprehensive & 6-DoF-track. Each track is further divided into simple and hard levels. The position-simple track includes tasks related to front & back & left & right spatial relationships, while the position-hard track includes tasks related to between, center, and customized. The orientation-simple track includes tasks related to the orientation of object parts, and the orientation-hard track includes tasks related to whether the object is upright or flipped (with very strict requirements for angles in both upright and flipped cases). Comprehensive tasks involve complex instruction understanding and long-horizon tasks; 6-DoF tasks simultaneously include requirements for both object position and orientation instructions. In Table 13, we present the complete task instructions, as well as the performance metrics of SoFar and the baseline. Due to the large number of tasks, we performed each task three times. It can be seen that SoFar achieves the best performance in all tracks, especially in the orientation-track and comprehensive & 6-DoF-track. We also show all the objects used in the real-world experiments in Fig. 17, covering a wide range of commonly and uncommonly used objects in daily life.

D.2 PointSO Model Details

For PointSO, we utilize FPS + KNN to perform patchify and employ a small PointNet [86] as the patch encoder. Subsequently, a standard Transformer encoder is adopted as the backbone, followed by a single linear layer to map the output to a three-dimensional vector space. All parameter configurations follow prior work on point cloud representation learning [26, 89, 91]. Detailed hyperparameter and model configurations are provided in Tables 14 and 15.

Table 15: **Details of PointSO model variants**. This table format follows Dosovitskiy et al. [28].

Model	CLIP	Layers	Hidden size	MLP size	Heads	#Params
Small	ViT-B/32	12	256	1024	4	11.4M
Base	ViT-B/32	12	384	1536	6	19.0M
Large	ViT-B/32	12	512	2048	8	43.6M

D.3 SoFar-LLaVA Model Details

In addition to leveraging the extensive knowledge and strong generalization capabilities of closed-source/open-source pretrained VLMs [98, 1, 3] for zero-shot or in-context learning, SoFar can also enhance the planning performance of open-source models through visual instruction tuning for rapid

Table 13: **Detailed zero-shot real-world 6-DoF rearrangement results**.

Task	CoPa [44]	ReKep-Auto [47]	SoFar-LLaVA (Ours)	SoFar (Ours)
Positional Object Manipulation				
Move the soccer ball to the right of the bread.	2/3	3/3	3/3	**3/3**
Place the doll to the right of the lemon.	3/3	3/3	3/3	**3/3**
Put the pliers on the right side of the soccer ball.	1/3	1/3	3/3	**2/3**
Move the pen to the right of the doll.	3/3	2/3	3/3	**3/3**
Place the carrot on the left of the croissant.	2/3	3/3	2/3	**2/3**
Move the avocado to the left of the baseball.	3/3	2/3	2/3	**3/3**
Pick the pepper and place it to the left of the charger.	1/3	2/3	2/3	**2/3**
Place the baseball on the left side of the mug.	3/3	2/3	2/3	**3/3**
Arrange the flower in front of the potato.	2/3	3/3	2/3	**3/3**
Put the volleyball in front of the knife.	3/3	3/3	3/3	**3/3**
Place the ice cream cone in front of the potato.	2/3	3/3	2/3	**3/3**
Move the bitter melon to the front of the forklift.	2/3	1/3	2/3	**2/3**
Place the orange at the back of the stapler.	3/3	2/3	3/3	**3/3**
Move the panda toy to the back of the shampoo bottle.	2/3	3/3	3/3	**2/3**
pick the pumpkin and place it behind the pomegranate.	**3/3**	2/3	1/3	2/3
Place the basketball at the back of the board wipe.	2/3	2/3	3/3	**2/3**
Put the apple inside the box.	3/3	2/3	3/3	**3/3**
Place the waffles on the center of the plate.	3/3	2/3	3/3	**3/3**
Move the hamburger into the bowl.	2/3	2/3	2/3	**3/3**
Pick the puppet and put it into the basket.	1/3	2/3	2/3	**2/3**
Drop the grape into the box.	2/3	3/3	3/3	**2/3**
Put the doll between the lemon and the USB.	2/3	2/3	2/3	**3/3**
Set the duck toy in the center of the cart, bowl, and camera.	2/3	1/3	2/3	**2/3**
Place the strawberry between the Coke bottle and the glue.	2/3	2/3	3/3	**3/3**
Put the pen behind the basketball and in front of the vase.	2/3	1/3	2/3	**2/3**
Total success rate	74.7%	72.0%	81.3%	**85.3%**
Orientational Object Manipulation				
Turn the yellow head of the toy car to the right.	2/3	2/3	1/3	**2/3**
Adjust the knife handle so it points to the right.	2/3	1/3	2/3	**2/3**
Rotate the cap of the bottle towards the right.	2/3	2/3	2/3	**2/3**
Rotate the tip of the screwdriver to face the right.	0/3	0/3	1/3	**1/3**
Rotate the stem of the apple to the right.	0/3	1/3	1/3	**2/3**
Turn the front of the toy car to the left.	0/3	0/3	2/3	**2/3**
Rotate the cap of the bottle towards the left.	2/3	1/3	1/3	**2/3**
Adjust the pear's stem to the right.	1/3	1/3	1/3	**1/3**
Turn the mug handle to the right.	1/3	1/3	2/3	**2/3**
Rotate the handle of the mug to towards right.	2/3	1/3	**2/3**	1/3
Rotate the box so the text side faces forward.	0/3	1/3	0/3	**1/3**
Adjust the USB port to point forward.	0/3	0/3	1/3	**1/3**
Set the bottle upright.	0/3	1/3	0/3	**1/3**
Place the coffee cup in an upright position.	1/3	1/3	2/3	**2/3**
Upright the statue of liberty	0/3	0/3	**1/3**	0/3
Stand the doll upright.	0/3	1/3	0/3	**1/3**
Right the Coke can.	0/3	0/3	1/3	**1/3**
Flip the bottle upside down.	0/3	0/3	0/3	**1/3**
Turn the coffee cup upside down.	0/3	0/3	1/3	**1/3**
Invert the shampoo bottle upside down.	0/3	0/3	0/3	**0/3**
Total success rate	21.7%	23.3%	35.0%	**43.3%**
Comprehensive 6-DoF Object Manipulation				
Pull out a tissue.	3/3	3/3	2/3	**3/3**
Place the right bottle into the box and arrange it in a 3×3 pattern.	0/3	0/3	0/3	**1/3**
Take the tallest box and position it on the right side.	1/3	1/3	3/3	**3/3**
Grasp the error bottle and put it on the right side.	1/3	2/3	1/3	**2/3**
Take out the green test tube and place it between the two bottles.	2/3	2/3	3/3	**3/3**
Pack the objects on the table into the box one by one.	1/3	1/3	0/3	**1/3**
Rotate the loopy doll to face the yellow dragon doll	0/3	1/3	1/3	**1/3**
Right the fallen wine glass and arrange it neatly in a row.	0/3	0/3	0/3	**0/3**
Grasp the handle of the knife and cut the bread.	0/3	0/3	0/3	**1/3**
Pick the baseball into the cart and turn the cart to facing right.	0/3	0/3	1/3	**2/3**
Place the mug on the left of the ball and the handle turn right.	0/3	0/3	1/3	**1/3**
Aim the camera at the toy truck.	1/3	0/3	1/3	**1/3**
Rotate the flashlight to illuminate the loopy.	0/3	0/3	1/3	**1/3**
Put the pen into the pen container.	0/3	1/3	0/3	**1/3**
Pour out chips from the chips cylinder to the plate.	0/3	1/3	1/3	**1/3**
Total success rate	20.0%	26.7%	33.3%	**48.9%**

fine-tuning. The pipeline of the model is illustrated in Fig. 18. A JSON-formatted 6-DoF scene graph, processed through a text tokenizer, along with the image refined by SoM [129], is fed into an LLM (*e.g.*, LLaMA [110, 111]) for supervised fine-tuning [67]. In the Open6DOR [25] task, we supplement the training dataset with additional samples retrieved and manually annotated from Objaverse [20], ensuring alignment with the object categories in the original benchmark. This dataset

Table 14: **Training recipes for PointSO and SoFar-LLaVA.**

	PointSO				SoFar-LLaVA
Config	Small	Base	Large	Finetune	SFT
optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
learning rate	5e-5	5e-5	2e-5	5e-5	2e-5
weight decay	5e-2	5e-2	5e-2	5e-2	0
learning rate scheduler	cosine	cosine	cosine	cosine	cosine
training epochs	300	300	300	50	2
warmup epochs	10	10	10	5	0.03
batch size	256	256	256	256	128
drop path rate	0.2	0.2	0.2	0.2	-
number of points	10000	10000	10000	10000	-
number of point patches	512	512	512	512	-
point patch size	32	32	32	32	-
augmentation	Rot&Part&Noise	Rot&Part&Noise	Rot&Part&Noise	Rotation	-
GPU device	8×H800	8×H800	8×H800	8×H800	8×H800

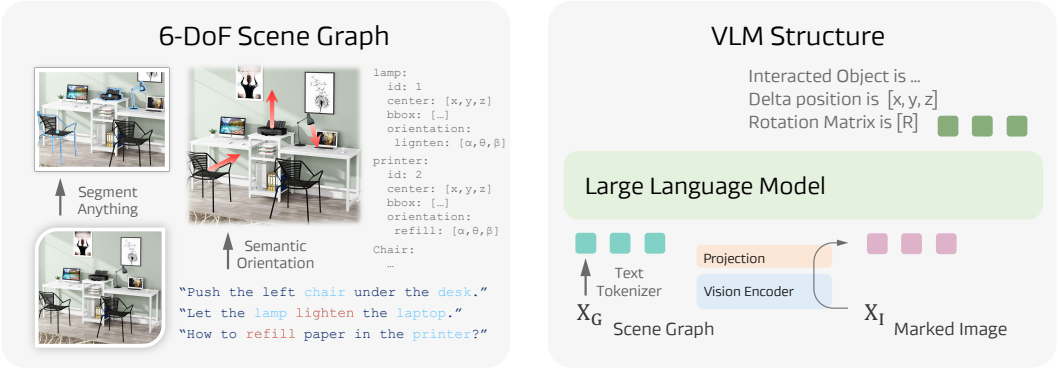


Figure 18: **Pipeline of SoFar-LLaVA**, a fine-tuned VLM based on visual instruction tuning.

includes approximately 3,000 6-DoF object manipulation instructions. Using this data, we construct dialogue-style training data based on ChatGPT and train the SoFar-LLaVA model. The training hyperparameters are detailed in Table 14. Similarly, we finetune PointSO on this training dataset and achieve superior performance on the Open6DOR task.

D.4 ChatGPT API Costs

The knowledge of OrienText300K is derived from the annotations of 3D modelers on Sketchfab, combined with ChatGPT's filtering and comprehension capabilities. To generate semantic orientation annotations, we filter the 800K dataset of Objaverse [20] and apply ChatGPT to approximately 350K of the filtered data to generate semantic text-view index pairs. The OpenAI official API was used for these calls, with the GPT-4o version set to 2024-08-06 and the output format configured as JSON. The total cost for debugging and execution amounted to approximately $10K.

E Additional Benchmark Statistic Analysis

E.1 6-DoF SpatialBench Analysis

We conduct a statistical analysis of the manually constructed 6-DoF SpatialBench, with category comparisons and word cloud visualizations shown in Fig. 19. We collect diverse image data from the internet, encompassing scenes such as indoor, outdoor, and natural landscapes. The questions may involve one or multiple objects, with varying levels of uncertainty in image resolution. Most importantly, we are the first to propose a VQA benchmark for orientation understanding, focusing on both quantitative and qualitative evaluation of orientation.

E.2 Open6DOR V2 Analysis

Open6DOR V2 builds upon Open6DOR V1 by removing some incorrectly labeled data, removing manual evaluation metrics, and integrating assets and metrics into Libero, enabling closed-loop policy evaluation. The detailed number of tasks is presented in Table 16, comprising over 4,500 tasks in total. Notably, we remove level 2 of the position track in Open6DOR V1 [25] because it requires manual inspection, which is not conducive to open-source use and replication by the community. Besides, due to the randomness of object drops in the scene, approximately 8% of the samples already satisfy the evaluation metrics in their initial state.

F Related Works

F.1 Vision-Language Models for Spatial Understanding

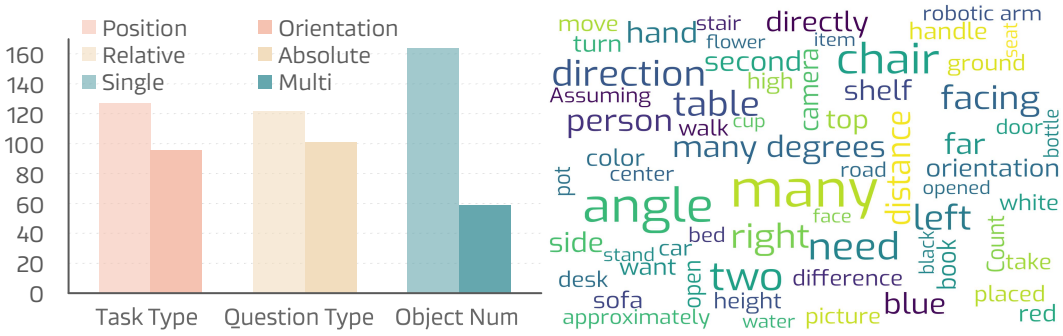
Vision-Language Models are rapidly being developed in the research community, driven by the storming lead in extending GPT-style [95, 96, 7] Large Language Models (LLMs) like LLaMA [110, 111] to VLMs [67, 68, 27, 105, 141, 144, 51]. SpatialVLM [10] pioneers this direction by constructing VQA data in spatial understanding from RGB-D, which is used for training an RGB-only VLM. Following SpatialVLM, SpatialRGPT [12] extends RGB-based spatial understanding to RGB-D by constructing spatial understanding data using 3D scene graphs. SpatialBot [8] explores RGB-D spatial reasoning through hierarchical depth-based reasoning. Some other works propose visual prompting for improving GPT-4V's spatial understanding [65, 129, 75]. Meanwhile, another line of works explores VLMs using 3D representations such as point clouds for 3D scene [43, 35] and object-centric [91, 127, 92] understanding. More recently, OmniSpatial [51] proposed a comprehensive and challenging spatial reasoning benchmark. Despite the remarkable progress, these works are limited to 3-DoF understanding, which is not actionable. In contrast, we explore spatial understanding in 6-DoFs from RGB-D via VLMs. Unlike vanilla 3D scene graphs used by SpatialRGPT for data construction, we propose orientation-aware 3D scene graphs realized by our proposed PointSO. In addition, we formulate spatial understanding as graph learning, where the scene graph nodes are directly input during inference.

F.2 Language-Grounded Robot Manipulation

Language-grounded robot Manipulation adopts the human language as a general instruction interface. Existing works can be categorized into two groups: **i)** *End-to-end* models like RT-series [6, 152, 4] built upon unified cross-modal Transformers with tokenized actions [100, 66, 145], large vision-language-action models built from VLMs [56, 143], or 3D representations [146, 137]. Training on robot data such as Open X-Embodiment [15] and DROID [55], a remarkable process has been made. However, the data *scale* is still limited compared to in-the-wild data for training VLMs. **ii)** *Decoupled* high-level reasoning and low-level actions in large VLMs and small off-the-shelf policy models, primitives [49, 63, 46, 44, 34, 131, 138, 77, 39, 149], or articulated priors [45, 60]. Our SoFar lies in this group, where an open-world generalization property emerges from VLMs and our proposed PointSO is empowered by orientation-aware spatial understanding.

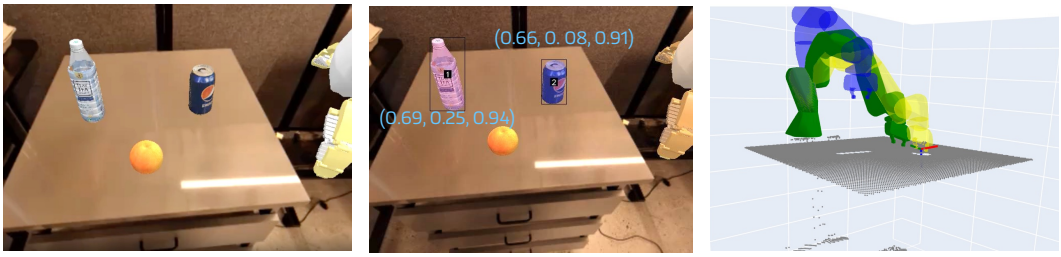
F.3 3D Representation Learning

Research on 3D Representation Learning encompasses various methods, including point-based [86, 87], voxel-based [78], and multiview-based approaches [102, 38]. Point-based methods [93, 31] have gained prominence in object classification [122, 113] due to their sparsity yet geometry-informative representation. On the other hand, voxel-based methods [21, 134, 90] offer dense representation and translation invariance, leading to a remarkable performance in object detection [17] and segmentation [133, 2]. The evolution of attention mechanisms [114] has also contributed to the development of effective representations for downstream tasks, as exemplified by the emergence of 3D Transformers [31, 74, 76]. Notably, 3D self-supervised representation learning has garnered significant attention in recent studies. PointContrast [126] utilizes contrastive learning across different views to acquire discriminative 3D scene representations. Innovations such as Point-BERT [136] and Point-MAE [83] introduce masked modeling [40, 22] pretraining into the 3D domain. ACT [26] pioneers cross-modal geometry understanding through 2D or language foundation models such as CLIP [97] or BERT [22]. Following ACT, ReCon [89] further proposes a learning paradigm



(a): Statistical Analysis (b): Word Cloud

Figure 19: **6-DoF SpatialBench statistics**. (a) Statistical analysis of the task type, question type, and object relation. (b) Word cloud visualization.



(a): Visual Observation (b): Scene Analysis (c): Motion Planning

Figure 20: An example of SoFar how to finish "move near" task in SIMPLER [62].

that unifies generative and contrastive learning. PPT [142] highlights the significance of positional encoding in 3D representation learning. Additionally, leveraging foundation vision-language models like CLIP [26, 97] has spurred the exploration of a new direction in open-world 3D representation learning. This line of work seeks to extend the applicability and adaptability of 3D representations in diverse and open-world/vocabulary scenarios [84, 140, 24, 23, 32].

G Additional Discussions

G.1 Relation to Affordance & 6-DoF Pose Estimation

Conceptually, this semantic orientation is a counterpart of *affordance* [36, 101, 88, 37] but beyond, as SO and affordance, all present potential actions and interactions with objects. However, SO also contains the spatial understanding of intra-object part-level attributes more than affordance learning. Compared to vanilla 6-DoF pose estimation, our proposed SO combined with the 3-DoF translation understanding, has the same DoF completeness. The difference is, our proposed SO is grounded in languages, making it useful for open-world manipulation requiring complicated spatial reasoning [109, 49, 25]. In addition, our Semantic Orientation can be auto-labeled from Internet 3D data that achieves higher scalability, as introduced in Section 2.

G.2 Comparison to Concurrent Works

Comparison with ReKep [47]

Recently, ReKep has succeeded in executing complex robotic tasks, such as long-horizon manipulation, based on the relationships and constraints between spatial key points. Its structural design offers many insights that SoFar can draw upon, yet it also presents several issues: (1) Overly customized prompt engineering. ReKep requires manually designed complex system prompts for each task during inference. While this approach may be described as "no training", it cannot be considered a true zero-shot transfer. In contrast, SoFar achieves genuine zero-shot transfer by eliminating the need for any human involvement during inference; (2) Using constraints based solely on key

Table 16: **Statistics of Open6DOR V2 Benchmark.** The entire benchmark comprises three independent tracks, each featuring diverse tasks with careful annotations. The tasks are divided into different levels based on instruction categories, with statistics demonstrated above.

Track	Position-track							Rotation-track			6-DoF-track	Totel
Level	Level 0					Level 1		Level 0	Level 1	Level 2	-	-
Task Catog.	Left	Right	Top	Behind	Front	Between	Center	Geometric	Directional	Semantic	-	-
Task Stat.	296	266	209	297	278	193	159	318	367	134	1810	4535
Benchmark Stat.	1698							1027			1810	4535

points fails to capture the full 6-DoF pose integrity of objects. For example, in the "pouring water" task, merely bringing the spout of the kettle close to the cup may lead to incorrect solutions, such as the kettle overturning; (3) ReKep requires all key points to be present in the first frame, and each step of the process—from mask extraction to feature dimensionality reduction, clustering, and filtering—introduces additional hyperparameters.

Comparison with Orient Anything [118]

Recently, Orient Anything also highlighted the importance of orientation in spatial perception and adopted a training data construction approach similar to Our PointSO. Our primary distinction lies in semantic orientation, which is language-conditioned orientation. In contrast, Orient Anything is limited to learning basic directions such as "front" and "top". By aligning with textual information, semantic orientation better enhances spatial perception, understanding, and robotic manipulation.

G.3 Future Works

Future work includes further expanding the OrienText300K with larger datasets like Objaverse-XL [19], enhancing the performance of semantic orientation through self-supervised learning and pretraining methods [40, 97, 26, 89], and demonstrating its effectiveness in a broader range of robotic scenarios, such as navigation [9], mobile manipulation [132], lifelong learning [64], spatio-temporal reasoning [47, 72, 73, 130], humanoid [41, 11, 14, 42], and human-robot interaction [70, 71].

H Additional Visualizations

H.1 Robotic Manipulation

As shown in Fig. 20, we present a visualization of executing a task named "move near". According to the input image and task instruction - "*move blue plastic bottle near pepsi can*", SoFar can predict the center coordinate of the target object (bottle) and relative target (pepsi can), and it would infer the place coordinate and produce a series of grasp poses.

H.2 6-DoF SpatialBench

To further evaluate 6-DoF spatial understanding, we construct a 6-DoF SpatialBench. We present examples of question-answer pairs from the 6-DoF SpatialBench, with quantitative and qualitative questions shown in Figs. 21 and 22, respectively. The benchmark we constructed is both challenging and practical, potentially involving calculations based on the laws of motion, such as "*Assuming a moving speed of 0.5 m/s, how many seconds would it take to walk from here to the white flower?*" Moreover, it covers a wide range of spatially relevant scenarios across both indoor and outdoor environments.

H.3 System Prompts

Prompt engineering significantly enhances ChatGPT's capabilities. The model's understanding and reasoning abilities can be greatly improved by leveraging techniques such as Chain-of-Thought [119] and In-Context Learning [7]. Figs. 23 and 24 illustrate the system prompt we used in constructing OrienText300K. Fig. 25, Fig. 26, and Fig. 27 illustrate the system prompt we used when evaluating SoFar on Open6DOR (simulation), object manipulation (both simulation and real worlds), and

VQA, respectively. Note that different from previous methods [46, 47], SoFar does not require complicated in-context examples.

I Broader impacts

Our work on semantic orientation significantly enhances robotic spatial reasoning and manipulation capabilities, enabling more intuitive human-robot interaction. This advancement can improve efficiency in various industries, such as manufacturing and healthcare, and enhance the quality of life by assisting in tasks like elderly care and home automation. Additionally, it contributes to the broader field of AI research by providing new tools and benchmarks for spatial reasoning and language-grounded manipulation.

[Task Type: Position Question Type: Absolute]

[**Question**]: Count from right to left and start at 1, which two of the red flower pots are the group of people in the middle of?

[A]: "4 and 5"
[B]: "2 and 3"
[C]: "1 and 2"
[D]: "3 and 4"

[Answer]: C

[Task Type: Orientation Question Type: Absolute]

[**Question**]: If you want to align the orientations of the two chairs, what is the minimum angle you need to rotate the chair on the right?

[A]: "75°"
[B]: "55°"
[C]: "35°"
[D]: "15°"

[Answer]: C

[Task Type: Position Question Type: Absolute]

[**Question**]: Assuming a moving speed of 0.5 m/s, how many seconds would it take to walk from here to the white flower?

[A]: "3s"
[B]: "5s"
[C]: "7s"
[D]: "10s"

[Answer]: B

[Task Type: Orientation Question Type: Absolute]

[**Question**]: How many white chairs are facing the window?

[A]: "2"
[B]: "4"
[C]: "1"
[D]: "3"

[Answer]: A

Figure 21: **Visualization example of 6-DoF SpatialBench data samples.**

[Task Type: Position Question Type: Relative]

[**Question**]: Which side of the steps is narrower?

[A]: "the left"
[B]: "the right"
[C]: "the middle"
[D]: "the same"

[Answer]: B

[Task Type: Orientation Question Type: Relative]

[**Question**]: Which direction does the handle of the cup in the upper right corner point to?

[A]: "left"
[B]: "right"
[C]: "front"
[D]: "back"

[Answer]: A

[Task Type: Position Question Type: Relative]

[**Question**]: How many compartments are there in the heart-shaped grid for storing books?

[A]: "5"
[B]: "3"
[C]: "6"
[D]: "4"

[Answer]: A

[Task Type: Orientation Question Type: Relative]

[**Question**]: If you are a driver driving a car on the road from near to far, which direction will you turn to?

[A]: "first turn left and then left"
[B]: "first turn right and then left"
[C]: "first turn left and then right"
[D]: "first turn right and then right"

[Answer]: C

Figure 22: **Visualization example of 6-DoF SpatialBench data samples.**

[System Prompt]

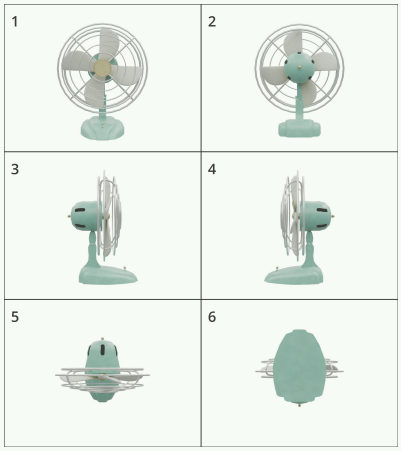
You are an expert AI assistant for 3D object understanding.
The user imported a potentially uncalibrated 3D model into Blender and placed cameras in front, back, left, right, top, and bottom to render images, labeled from 1 to 6.
You are required to infer the entire 3D object based on these images and determine its attributes.

Your task is to assess the following attributes for each 3D model and respond with "true" or "false" for each question:
Axis Alignment: Determine whether the object is horizontally and vertically aligned across all views. Key features (e.g., edges, handles, or other distinct elements) of the object must be perpendicular or parallel to the cameras. Respond "true" if all views are aligned with the axis, "false" if not.
Scene or Collection: Determine whether the 3D model represents a 3D scene or a collection of independent objects (e.g. a room, outdoor scene, or multiple independent objects). Respond with "true" if it does, and "false" if it only contains a single object.
White: Determine whether the 3D model only has single white or gray colors, and lacks any other colors. Respond with "true" if it is white or gray, and "false" if it has any other colors (e.g., black or yellow).
Ground: Determine whether the 3D model includes a ground plane for auxiliary visualization. Respond with "true" if it does, and "false" if it only has the object.
High Quality: Determine whether the 3D model is a refined, well-constructed mesh without defects, such as point noise or streaking artifacts commonly found in low-quality RGBD scans. Respond with "true" if the mesh is clean and smooth, and "false" if it contains noise, roughness, or visual artifacts.
Distinguishable Views: Determine whether the 3D model has distinguishable views, or has clear semantic information in certain views (e.g., some 3D object has clear front, top directions). Respond with "true" if the 6 views show noticeable differences or have clear semantic information in certain views, and "false" if the views appear identical and there is no obvious semantic information on all views.
Reasonable Object: Determine whether the 3D model represents a common, recognizable, meaningful object. Respond with "true" if it is, and "false" if it is abstract, confused, or unrecognizable.

You need to first analyze the 3D object detail, and then output its correct attributes.

[User]

Standard Views:



Oblique Views: (Only for reference)

Figure 23: **The system prompt of GPT-4o used for filtering Objaverse data.**

[System Prompt]

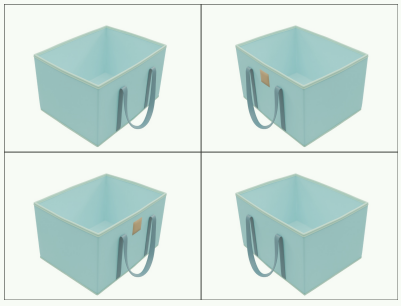
You are a visual assistant specializing in interpreting 3D objects from multiple perspectives.
You will receive 6 images of a 3D object from standard views (front, back, left, right, top, bottom), presented in random order. Typically, image 5 corresponds to the top view.
Your task is to generate an instruction-index pair that identifies a meaningful semantic orientation for the object, based on its function or commonly understood orientation.
The instruction can be a verb, noun, adjective, or phrase, and must clearly relate to the object's function or orientation in everyday use.
Ensure the direction is clear, objective, and uniquely meaningful.

Examples:
For a pen, the instruction might be "pen cap", and the index is the image with the pen cap facing the camera.
For a cup, the instruction might be "handle", and the index is the image with the cup handle facing the camera.
For a phone, the instruction might be "screen", and the index is the image with the phone screen facing the camera.
For a table, the instruction might be "on", and the index is the image with the tabletop facing the camera.
For a power outlet, the instruction might be "plug-in". Based on common knowledge, its semantic orientation is perpendicular to the power outlet's plane, along the direction of the power outlet's slots, and therefore, the index is the image with the power outlet pinholes' plane facing the camera.
For a desk, the instruction might be “open the drawer”. Based on common sense, the robot would need to pull the drawer open. The semantic orientation corresponds to the direction of the drawer's extension, hence the index is the image with the drawer handle facing the camera.
For a microphone, the instruction might be "speak", the semantic orientation is along the direction of the microphone's head, and therefore, the index is the image with the microphone head facing the camera.

You need to first analyze the category, attributes, characteristics, state, and usage of this 3D object in detail, and then output a pair of instructions and index.
When it is challenging to generate complex instructions, or when multiple views of the object are too similar to produce a unique instruction, you can use simpler instructions, such as "top" or "front".
The output format is as follows:
Analysis: "..."
Instruction: "..."
Index: 1-6

[User]

Oblique Views: (Only for reference)



Standard Views:

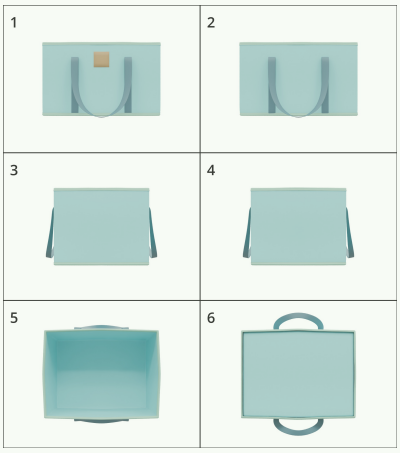


Figure 24: **The system prompt of GPT-4o used for generating semantic orientation-Index pairs.**

[Parsing System Prompt]

You are an assistant specialized in interpreting tabletop pick-and-place instructions for robotic manipulation.
Your main goals are to identify relevant objects and analyze necessary orientations.

Key Objectives
1. Object Identification: Identify and list the objects mentioned in the instruction. Exclude the table itself.
2. Orientation Analysis: For the object needs to pick & place, determine any required orientation crucial to the task's success. If orientation isn't specified, leave the orientation list empty.
3. Direction Terms: Limit directional terms to these two categories:
   - Object Parts: e.g., "handle", "pen cap", "top"
   - Interaction Actions: e.g., "pour out", "open"
   Terms must be single words, not phrases or sentences.
   You must analysis both the instruction and the image to determine the object's direction attributes.
4. Disambiguation of Identification: If instructions reference vague objects (e.g., "else object", "all objects"), use visual information to clarify.
5. Disambiguation of Orientation: If the instructions describe complex rotation like "upright", you can interpret them as ensuring an object's relevant part is aligned with the z-axis (e.g., "bottle cap", "top").
   This disambiguation utilizes world knowledge, as we define the far-to-near direction as the x-axis, the left-to-right direction as the y-axis, and the bottom-to-top direction as the z-axis.
   Similarly, place an object to point forward means that the "top" of the object is oriented along the x-axis.

[Reasoning System Prompt]

You are an assistant for spatial intelligence and robotic operations, specializing in pick-and-place tasks.
Your role is to process robotic commands to pick a object and place it in a specific location.

Input Context:
1. Pick & Place Command: A directive specifying which object to pick and where to place it, including any specific pose requirements.
2. picked object info: A dictionary with the picked object's position in the world coordinate system.
   - Coordinates: Object center and bounding box in 3D (x, y, z), where:
      -- x: Extends from far to near. Objects closer to the observer have larger x-values
      -- y: Extends from left to right. Objects further to the right have larger y-values
      -- z: Extends upward. Objects positioned higher have larger z-values
3. other objects info: A list of dictionaries with the position of other objects in the scene.
   - Coordinates: Object center and bounding box in 3D (x, y, z), same in the world coordinate system.

Objective:
1. Generate target placement position: Based on the spatial location descriptions provided in the instructions (e.g., 'behind,' 'between,' 'left,' etc.), as well as each object's center and bounding box (bbox), analyze and calculate the appropriate placement for the picked object.
   - front indicates positioning the object at an x-coordinate slightly larger than the reference object's x maximum.
   - right indicates positioning the object at a y-coordinate slightly larger than the reference object's y maximum.
   - between indicates positioning the object at the midpoint between two reference objects.

[User]

Place the knife behind the clipboard on the table.
And rotate the handle of the knife to left.

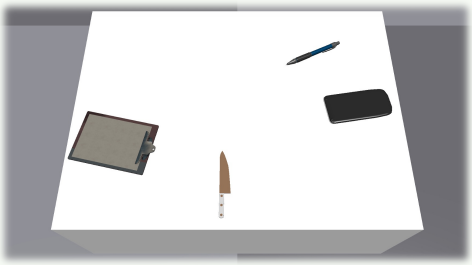


Figure 25: **The system prompt of Open6DOR tasks.**

[Parsing System Prompt]

You are a spatially intelligent AI specializing in interpreting objects, spatial directions, and interaction semantics for tasks involving spatial understanding or robotic manipulation.
The user will input an image and an instruction. Analyze user instruction and provide:

Objects: List involved objects using concise nouns or phrases, without any adjectives (e.g, the "top drawer" should be listed as "drawer").
Semantic Orientations: Identify essential spatial or action-related terms, categorized as:
- Object Parts: e.g., "handle", "lid", "top".
- Action Terms: e.g., "pour out", "open".

Guidelines:
Focus on key spatial or action contexts for task completion.
Use implicit spatial conventions (Certain user instructions need to satisfy implicit constraints related to position and orientation.) if practical.
Avoid numeric values or absolute positions.
Only specify object-centric pose relationships, not inter-object positions (such as left, right, front, behind).

[Reasoning System Prompt]

You are a robotic spatial intelligence and manipulation assistant, specialized in interpreting commands and scene structures for robotic object manipulation.
Your task is to analyze the user's directive and scene graph to guide the robot in identifying objects, computing spatial transformations, and producing step-by-step guidance for manipulation tasks.

Input Context:
1. Manipulation Command: A directive specifying which object to pick and where to place it, including any specific pose requirements.
2. Scene Graph: A dictionary with the scene objects' position and orientation in the world coordinate system.
   - Coordinates: Object center and bounding box in 3D (x, y, z), where:
     -- x: Extends from near to far. Objects further to the observer have larger x-values
     -- y: Extends from right to left. Objects further to the left have larger y-values
     -- z: Extends upward. Objects positioned higher have larger z-values
   - Orientations of the object's parts (e.g., 'screen', 'handle') in 3D space.
     -- (1, 0, 0): Points forward along the x-axis
     -- (0, 1, 0): Points left along the y-axis
     -- (0, 0, 1): Points upward along the z-axis

Objective: To process each command, follow these steps:

Target Identification: Identify the object to be picked up or manipulated.
Final Position: Specify the intended final position of the object after manipulation, in terms of x, y, z coordinates.
Orientation Mapping: For each semantic orientation provided, compute the final orientation of the manipulated object in the world coordinate system.

[User]

Open top drawer.

Figure 26: **The system prompt of general manipulation tasks.**

[Parsing System Prompt]

You are a spatially intelligent, embodied AI brain specialized in spatial and interactive understanding, tasked with interpreting objects, spatial directions, and relevant interaction semantics in response to the user's queries. The user provides commands or questions related to spatial intelligence or robotic manipulation, often with an image input. Your job is to analyze the given instruction and provide a list of objects involved in the task, alongside semantic orientations needed to complete the instruction effectively. You should focus on the key interaction directions required for successful completion without specifying numeric values or absolute positions, as these will be calculated by an expert model later.

Guidelines:
1. Focus on Semantic Orientations: Define directions concisely using single terms that fall into one of these two categories:
   - Object Parts (e.g., "handle", "screen", "top")
   - Action-Oriented Terms (e.g., "pour out", "plug-in", "open")
2. Optimize for Simplicity: Choose terms that provide essential spatial or action context while remaining simple and intuitive for the model. Use only the most relevant directions or parts needed to complete the user's task.
3. Analysis: When necessary, use implicit spatial conventions where appropriate to ensure a practical output for the model.
4. Only object-centric pose related: Distinguish which object relationships are determined by position (such as left, right, front, behind) and which are determined by object pose, and we only focus on the direction of object - centric pose.

[Reasoning System Prompt]

You are a spatial intelligence assistant specialized in understanding 3D visual scenes and answering spatial reasoning questions.

The user will provide:
Image: An image of the scene.
Question: User question about the spatial relationships between objects in the scene.
Scene Graph: Additional information about the objects, including:
   - id: object ID
   - object name: object category
   - center: 3D coordinates of the object's center
   - bounding box: 3D bounding box coordinates
   - orientation: object directions in 3D space

All the coordinates are in the camera coordinate system, where:
   - x-axis: Extends from left to right in the image, objects located right have larger x-values
   - y-axis: Extends from bottom to top in the image, objects located at top of the image have larger y-values
   - z-axis: Extends from near to far in the image, objects located further away have larger z-values

You need to focus mainly on the image, the scene graph information is just for reference.
Avoid providing answers such as "cannot determine." Instead, provide the most likely answer based on the information available.

[User]

How far between the left bottle and the right bottle?

Figure 27: **The system prompt of visual-question-answering tasks.**