2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Polaris: Open-ended Interactive Robotic Manipulation via Syn2Real Visual Grounding and Large Language Models

Tianyu Wang[1], Haitao Lin[1], Junqiu Yu[1], Yanwei Fu[1†]

***Abstract*— This paper investigates the task of the open-ended interactive robotic manipulation on table-top scenarios. While recent Large Language Models (LLMs) enhance robots' comprehension of user instructions, their lack of visual grounding constrains their ability to physically interact with the environment. This is because the robot needs to locate the target object for manipulation within the physical workspace. To this end, we introduce an interactive robotic manipulation framework called Polaris, which integrates perception and interaction by utilizing GPT-4 alongside grounded vision models. For precise manipulation, it is essential that such grounded vision models produce detailed object pose for the target object, rather than merely identifying pixels belonging to them in the image. Consequently, we propose a novel Synthetic-to-Real (Syn2Real) pose estimation pipeline. This pipeline utilizes rendered synthetic data for training and is then transferred to real-world manipulation tasks. The real-world performance demonstrates the efficacy of our proposed pipeline and underscores its potential for extension to more general categories. Moreover, real-robot experiments have showcased the impressive performance of our framework in grasping and executing multiple manipulation tasks. This indicates its potential to generalize to scenarios beyond the tabletop. More information and video results are available here: https://star-uu-wang.github.io/Polaris/.**

I. Introduction

The longstanding goal of robotics research has been to bridge the interaction between robots and humans for real-world grasping [1], [2] and manipulation tasks [3], [4]. Natural language instructions play a central role in open-ended human-robot interaction, guiding robots to accomplish various tasks [5], [6], [2], [7]. Recently, Large Language Models (LLMs) and Vision Language Models (VLMs) have made significant progress [8]. They possess extensive world knowledge and have demonstrated strong abilities to understand human instructions, leading to the development of numerous methods for translating language and visual inputs into robotic manipulation actions [9], [10], [11], [12]. These methods, with their diverse attempts across different dimensions of robotics research, prompt further consideration on how to fully leverage the perceptual and interactive capabilities of LLMs to support various robotic manipulations.

We explore the issue of open-ended interactive robotic manipulation on tabletop-level scenarios, such as "Please help me tidy the table". Previous studies [13], [14], [15] have attempted to tackle such challenges by employing LLMs as task planners, translating high-level instructions into action sequences they comprehend. However, existing methods for real-world robotic manipulation tasks often lack robustness in visual grounding and tend to overlook object affordances and action feasibility. Vision-centric robotic manipulation equips robots with environmental perceptual abilities, enabling action planning based on perception. Nevertheless, this necessitates high-quality, real-world annotated data.

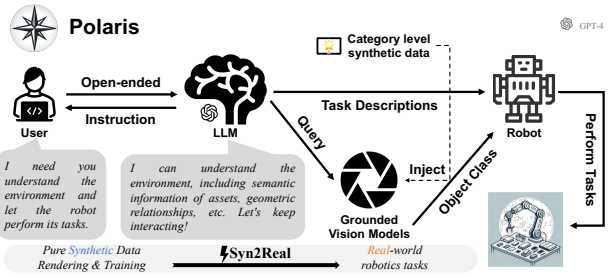


Fig. 1. **Polaris:** A tabletop-level object robotic manipulation framework centered on syn2real visual grounding driven by open-ended interaction with GPT-4. Users engage in continuous, open-ended interaction with LLM, which maintains an ongoing comprehension of the scenes. 3D synthetic data is integrated into the training of grounded vision modules to facilitate the execution of real-world tabletop-level robotic tasks.

To tackle the challenge of open-ended interactive robotic manipulation, we use the readily available and powerful Large Language Model (LLM)—GPT-4 [16] to comprehend and extract the target query from the user's intricate description. Once the target query for an object is established, the subsequent step involves the robot locating and grasping the object. Visual grounding enables agents to interpret the visual environment based on these queries, thus aiding in more intricate tasks and interactions [17]. Additionally, the 6 Degree-of-Freedom (DoF) object pose estimation serves as a basis for accurate manipulation [18]. Hence, it is imperative to develop a grounded vision module combined with a pose estimation model to obtain object poses for subsequent motion planning. A recent method [18] has extended pose estimation from instance-level to category-level and introduced a category-level dataset with pose annotations. However, this dataset only includes a limited number of categories. To encompass a broader range of categories, we propose an efficient pipeline for generating synthetic data. Leveraging off-the-shelf rendering technologies, we can produce synthetic images of objects with pose annotations. The purely synthetic data generated through rendering is utilized to train the category-level pose estimation model and conduct inference in real-world scenes, representing

† indicates corresponding author.

[1]Tianyu Wang, Haitao Lin, Junqiu Yu and Yanwei Fu are with Fudan University, China. Corresponding author's email: yanweifu@fudan.edu.cn.

[1]The computations in this research were performed using the CFFF platform of Fudan University. This work was supported in part by Shanghai Platform for Neuromorphic and AI Chip under Grant 17DZ2260900 (NeuHelium).

a novel Synthetic-to-Real (Syn2Real) approach. Ultimately, we seamlessly integrate the vision grounding module with the LLM and the robot planner, establishing an open-ended interactive robot framework.

Our framework, named **Polaris**, features syn2real visual grounding driven by GPT-4 to enhance tabletop-level interactive robotic manipulation, as depicted in Fig. 1. Specifically, the framework relies on LLM for scene perception and open-ended human-robot interaction. It trains the pose estimation model within the grounded vision module using purely synthetic data, interprets queries provided by the LLM, and ultimately executes tabletop-level tasks through a 6D pose-based planner, enabling continuous interaction.

Our contributions can be summarized as follows: (1) We have introduced an automated method for generating depth images and pose annotations when 3D models are available, leveraging a lightweight rendering engine. Additionally, we have trained MVPoseNet6D using synthetic data and evaluated the model on real-world images. The results indicate that our method achieves syn2real category-level pose estimation and can be readily expanded to cover a wider range of categories. (2) Building upon syn2real visual grounding and GPT-4, we have proposed a novel framework called Polaris to address the challenge of open-ended interactive robotic manipulation. (3) We demonstrated Polaris's capabilities through real-robot grasping and manipulation experiments, showcasing efficient interaction, operational effectiveness, and satisfactory success rates across various tasks.

II. Related Work

LLMs for Robotics. Embodied intelligence mainly focuses on building systems where agents can purposefully exchange energy and information with the physical environment. It requires a correct understanding of the embodied perception process from a high-dimensional cognitive perspective to a low-dimensional execution perspective [19], [20]. Recent work [9] has shown that using LLMs as robotic brain can unify egocentric memory and control by studying downstream tasks of active exploration and embodied question answering. However, such new framework's perception system has flaws in its visual grounding, hindering robot-environment interaction, which will be addressed in this paper. On the other hand, there are zero-shot or few-shot methods [21], [14], [12], [13], [11], [15] that utilize LLMs as task planners, decomposing high-level instructions into executable primitive tasks. These methods assume the ability of robots to execute advanced commands. Unfortunately, they have not yet fully supported the open-ended interaction with robot and not robust enough due to insufficient perception of environment. Instead, our framework addresses these issues, providing a flexible paradigm that bridges users, LLMs, and robots, offering a new perspective on universal human-robot interaction.

Category-level Object Pose Estimation. The 6D object pose estimation is crucial in various applications, such as robotic manipulation and autonomous driving. The objective of category-level object pose estimation is to predict the 6D pose and 3D size of diverse instances belonging to a shared category. The current mainstream methods can be divided into two types: RGB-D based and depth based only. RGB-D based methods [18], [22], [23], [24], [25] often leverages color cues for improved object recognition, which can capture fine-grained texture details, enhancing feature extraction. However, RGB-D based methods often encounter some challenges, such as being sensitive to lighting conditions and color variations. Depth based methods [26], [27], [28], [29], [30] rely solely on depth information, which reduce data complexity and lead to faster processing potentially. The above methods often involve scanning real objects or annotating images of real scenes. Based on the recognition that SAR-Net [26] is depth based only and the affordance of synthetic data, we opt to further enhance the category-level pose estimation capabilities of SAR-Net and realize real-world application via synthetic-to-real. Given the framework demands for operational efficiency, our work aims to support a large scale of categories with a minimal number of parameters, thereby establishing a lightweight and easily expandable category-level data rendering and training architecture.

Vision-centric Interactive Robot Manipulation. In the realm of vision-centric interactive robot manipulation, recent advancements focus on enhancing robots' ability to perform tasks by learning from human demonstrations and integrating LLMs or Large Multimodal Models (LMMs) for better understanding and execution of vision-centric tasks. Interactive robot manipulation learning from human demonstrations frequently demands high-quality human videos or teleoperation data [31], [32], [33], [34], [35], [36], [37]. Simultaneously, it requires dependable reinforcement learning or imitation learning algorithms for the training of robot policies [38], [39], [40], [41], [42], [43]. While these methods offer considerable flexibility, they all necessitate the collection of human demonstrations through various means to learn different tasks, often requiring real-world physical annotations [44], [45]. Interactive robotic manipulation frequently requires affordance learning for objects based on visual inputs, where zero-shot [10], [46], few-shot [47], [48], and open-ended learning [49], [50], [51] are of significant interest. Open-ended learning methods facilitate the update and expansion of category sets and also provide a broader interaction space for human-in-loop tasks. Our proposed model employs a convenient and efficient method to render synthetic data for training the pose estimation models within the grounded vision module. By integrating with GPT-4, it addresses object affordance and supports open-ended human-robot interaction.

III. Problem Formulation

We present a novel open-ended interactive robotic manipulation problem via syn2real visual grounding and LLMs here, which shall have the following desired properties.

Property 1. 3D synthetic data rendering: For arbitrary 3D model assets, categorized by object class, synthetic instance data is needed to be collected from various viewpoints through a virtual engine and then injected into the training of subsequent category-level pose estimation model.

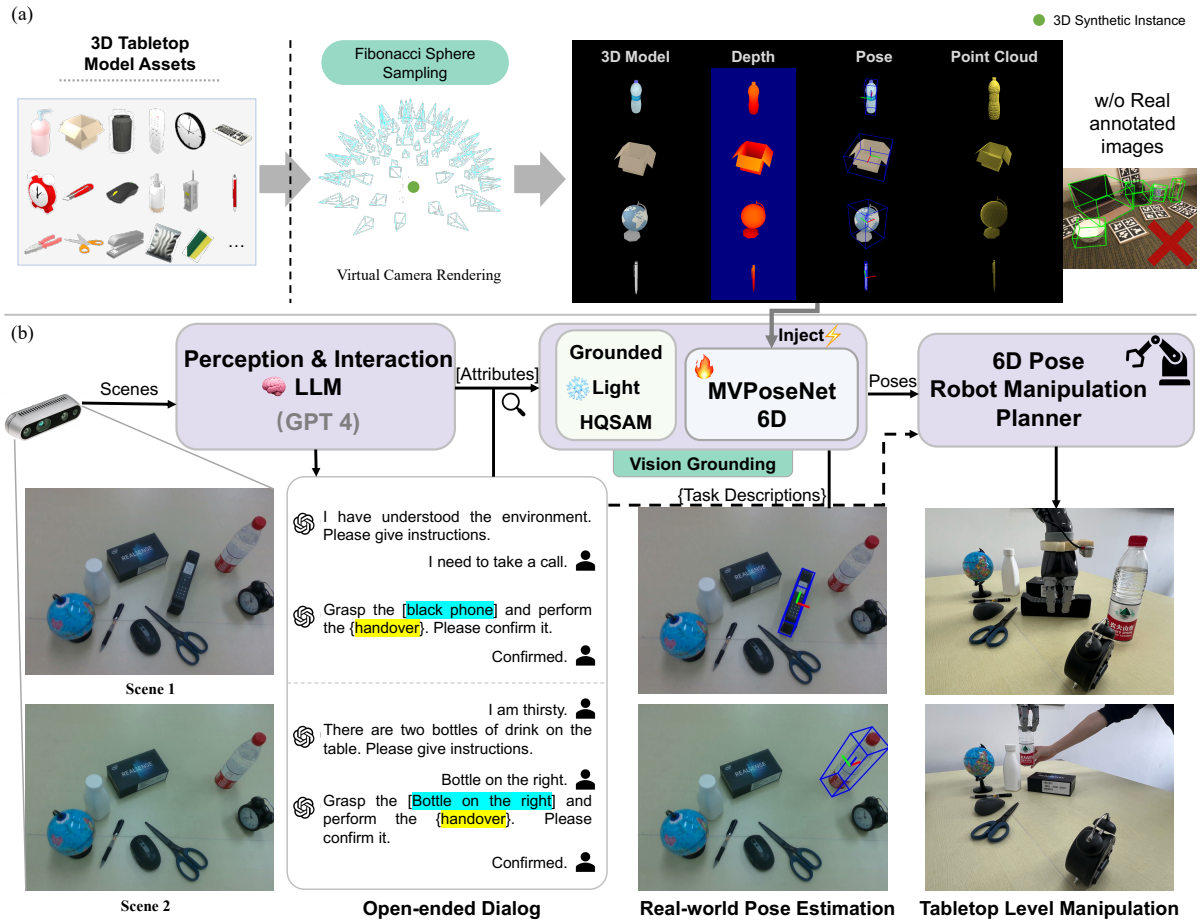


Fig. 2. **Overview of our framework.** (a) 3D synthetic data rendering. During rendering, we automatically generate various synthetic data by loading 3D model assets into a simulation engine and deploying dynamic virtual camera. We use the Fibonacci Sphere Sampling to select rendering viewpoints, to generate corresponding RGB, depth, pose, and observable point clouds. (b) The vision-centric robotic task pipeline. Given the image of the scene, which GPT-4, prompted as a scene perception and interaction LLM, interprets to understand instructions and describe objects and tasks. Our parser interprets these descriptions. We freeze the pre-trained detector and segmentation model within the grounded vision models and use a synthetic dataset to train the category-level pose estimation model. After retrieving object attributes, the model predicts poses based on the scene, allowing a 6D pose robot manipulation planner to execute real-world tasks.

Property 2. Open-ended interactive robotic manipulation: Given the RGB-D images of the scenes, the prompted LLM must comprehend bothe the scene and the user's language instructions, labelling the target objects for interaction and specific task descriptions. The robot must grasp and execute tasks according to the 6D pose of the target objects.

In terms of above descriptions, we give the typical setups to validate our framework: **(i)** We are particularly interested in table-level robotic tasks, where the variety of reachable objects is confined within an almost known domain. Hence, we opt to render common table objects category data for training the pose estimation model. However, based on our proposed rendering method, it is feasible to collect additional synthetic pose data from existing datasets ([52], [53], etc) or custom 3D modeling data, which can be generalized to the training of pose estimation models with greater capacity. **(ii)** Furthermore, we prefer the manner of interactive instructions. Specifically, user natural language instructions may not directly specify the objects for interaction (some queries may even be ambiguous), necessitating understanding and labelling by the LLMs. Additionally, continuous interaction is required, with the LLMs needing to keep up with scene changes and next user instructions.

IV. Method

Polaris is a sophisticated interactive robotic manipulation framework integrating perception and interaction, employing a LLM, specifically GPT-4, with grounded vision models. An overview of the proposed Polaris framework is presented in Fig. 2. In the ensuing subsections, we will detail the synthetic data rendering (Sec. IV-A), synthetic-to-real category-level pose estimation (Sec. IV-B, and open-ended interactive robotic manipulation design (Sec. IV-C).

A. 3D Synthetic Data Rendering

Given a 3D model M, we aim to render the RGB images I, depth image D, partial point cloud P and calculate the 6-DoF pose transformation $T = (R, t)$ and 3D size s of the model from current camera viewpoints.

Leveraging the SAPIEN [54] simulation environment, we utilize a subset of 3D models from the PartNet-Mobility dataset [53], supplemented with custom CAD modeling

data. To acquire the rendered images for each 3D model, we position the model at the origin of the world frame, variously adjusting the camera viewpoint to capture and render corresponding depth images. The PartNet-Mobility dataset comprises 2,000 articulated objects with motion annotations and rendering materials. This dataset serves as a valuable resource for advancing research in generalizable computer vision and manipulation, representing a continuation of the pioneering work in ShapeNet and PartNet.

In particular, to capture a broader range of camera viewpoints, we employ Fibonacci sphere sampling method to evenly distribute the camera positions across a sphere, as shown in Fig. 2 (a). Additionally, we introduce random in-plane rotations to each camera's orientation, expanding the coverage to encompass a more diverse set of camera angles.

Ultimately, we rendered a total of 24 tabletop-level object classes, including {"Bottle", "Box", "Dispenser", "Remote", "Camera", "Clock", "Eyeglasses", "Fan", "Faucet", "Globe", "Kettle", "Keyboard", "Knife", "Lamp", "Laptop", "Mouse", "Pen", "Phone", "Pliers", "Scissors", "Stapler", "USB", "Packaging", "Sponge"}, with 1K instances, resulting in 300K depth images along with poses, as illustrated in Fig. 2 (a). Additionally, the corresponding RGB and point cloud were generated simultaneously. This stage of the process solely relied on the CPU, making it very efficient. The pseudocode for the rendering process is provided by Algorithm 1.

Algorithm 1: Synthetic Data Rendering

Input: 3D models I containing N categories
Output: Fibonacci sphere rendering data for instances
Initialization: Set the rendering engine and parameters
for $i \leftarrow 1$ **to** N **do**
  The number of instances W contained in class i;
  for $j \leftarrow 1$ **to** W **do**
    Load the URDF model U_j of instance j;
    $(X_{min}, Y_{min}, Z_{min}) \leftarrow \infty$, $(X_{max}, Y_{max}, Z_{max}) \leftarrow -\infty$
    The number of parts S contained in model U_j;
    for $k \leftarrow 1$ **to** S **do**
      Load the points P_j^k of the part k;
      $(x_{min}, y_{min}, z_{min}) \leftarrow P_{j\ min}^k$, $(x_{max}, y_{max}, z_{max}) \leftarrow P_{j\ max}^k$
      Update global extreme point $(X_{min}, Y_{min}, Z_{min})$ and $(X_{max}, Y_{max}, Z_{max})$;
    Compute scale S_j by $(S_j^X, S_j^Y, S_j^Z) \leftarrow (X_{max} - X_{min}, Y_{max} - Y_{min}, Z_{max} - Z_{min})$;
    Generate camera poses τ by *Sphere Sampling*;
    for $n, \tau_n \leftarrow enumerate(\tau)$ **do**
      Mount dynamic virtual camera τ_n;
      Get instance pose $\lambda_j^n \leftarrow {\tau_n}^{-1}$;
      Update render to get RGB, PointCloud and Depth under τ_n;

B. Syn2Real Category-level Pose Estimation

Considering the efficient runtime requirements of the robotic manipulation, we aim to support a greater number of object categories with a smaller number of parameters and to robustly facilitate pose estimation in tabletop scenarios under various lighting conditions. We extend the output dimension of the original decoder in the depth-only SAR-Net [26] from 6 to 24 to accommodate the 24 new categories. By utilizing synthetically rendered multi-view data, the training process remains consistent with the original SAR-Net. This enables the model to learn shared geometric features among intra-category instances from different views of observed shapes. For the processing of category-level templates, we randomly select a general instance within the class, perform Object Canonicalization to align the coordinate system, and then execute Poisson Sampling and Farthest Point Sampling (FPS) to extract the category-level template point cloud. We transfer the model trained on synthetic data to the inference module, to support real-world category-level object pose estimation.

C. Open-ended Interactive Robotic Manipulation

As shown in Fig. 2 (b), our open-ended robot interaction framework primarily comprises three modules: a LLM that supports scene perception and human-robot interaction, a vision grounding module, and a robotic manipulation planner based on the 6D pose of objects.

Perception and Interaction LLM. Based on scene inputs from a depth camera, the framework we aim to construct should be capable of perceiving the scene, identifying object assets on the tabletop, and engaging in continuous interaction with the user based on the affordance of these assets and user requirements. By leveraging GPT-4's capabilities in image understanding, semantic extraction, and its powerful ability to comprehend user instructions, we call the GPT-4 API and prompt it to serve as the high-level perception and interaction brain for the robot. Firstly, we provide GPT-4 with a system-level explanation. This explanation is designed to guide the LLM to affirm its role and confine it within a specific domain, ensuring robustness and professionalism in task analysis during robotic manipulation. The primary task of the LLM before interacting with users is to understand and learn about the spatial specifics of tasks, rather than initiating work directly. Simultaneously, we expect the responses from the LLM to be task-oriented, necessitating specific object queries and task descriptions. We have constructed a parser that, upon the LLM's comprehension of the user's intent during the i round of interaction and subsequent user confirmation of the robotic instructions, extracts the object attributes A_i and task descriptions T_i from the instructions. These are then passed on to the vision grounding module and the robot planner.

Visual Grounding. The vision grounding module, serving as the core of vision-centric robotic manipulation, receives raw RGB images of the scene captured by depth cameras, depth information, and attributes A_i derived from the parsing of instruction provided by the LLM. We treat the attributes A_i as a query, which is then sent to the frozen Grounded-Light-HQSAM. Grounded-Light-HQSAM is an integrated model that incorporates Grounding DINO [55] and HQ-SAM [56]. Grounding DINO functions as an open-set object detector, utilizing visual-language modality fusion to generate bounding boxes and labels with free-form referring expressions. This process involves multiple phases, including a feature enhancer, a language-guided query selection module, and a cross-modality decoder. Once the grounding box of the target object

is obtained, it is used as a prompt for the segmentation model. Grounded-Light-HQSAM is capable of generating refined object masks in a lightweight and relatively fast manner. These masks are then used to crop the depth pixels belonging to the object, facilitating follow up pose and size estimation. After obtaining the depth, grounding mask and category label of the target object, we use the trained MVPoseNet6D to recover the real-world object 6-DoF pose and 3D size in the camera coordinate system, which provides information about the object's state in the current scene. Then, the estimated pose is transformed into the robot's base coordinate system for further motion planning.

6D Pose Robot Manipulation Planner. First of all, we need to answer a question: *Why is it necessary to consider the object's pose instead of using a straightforward grasp pose estimation method?* - While methods for direct 6-DoF grasp pose estimation [57], [58], [59] have made significant progress, their scope remains limited and is primarily applicable to pick-and-place operations. In our framework, we integrate pose estimation methods because the state of objects in 3D space is crucial for calculating meaningful manipulation points in various operational tasks, such as pouring water, handover, and some compositional tasks. We believe that object pose provides robots with rich contextual knowledge before proceeding with the motion planning. To ensure the extensibility of Polaris, we follow the principle of first inferring the object's pose and then calculating useful target gripper poses for precise manipulations. Particularly, as the pose of intra-category instances are pre-canonicalized [18], it becomes advantageous to define category-level grasp poses relevant to each task. These defined grasp poses are then transformed from object coordinates to camera coordinates using the estimated object 6D pose. The robot executes motion planning to move its gripper to the target grasp pose to complete each task. Thus, we have constructed a task-oriented 6D Pose Robot Manipulation Planner.

V. Experiment

We conducted a series of experiments, which included evaluating the synthetic-to-real pose estimation in both single-object and multi-object real-world scenarios, as well as testing the proposed Polaris framework against several baseline methods. The goals of the experiments are 1) to investigate the feasibility of applying MVPoseNet6D, trained purely on synthetic data, in real-world applications; 2) to demonstrate that our Polaris can efficiently achieve elaborate human-robot interaction in various scenarios; 3) to show the accuracy of our tabletop-level robotic manipulation system.

Polaris was deployed on a PC workstation with an Intel i9-13900K CPU and an NVIDIA RTX 6000 Ada Generation GPU. We used a KINOVA GEN2 robot with a Realsense D435 depth camera mounted in an eye-in-hand configuration. The tabletop-level testing objects are shown in Fig. 3.

A. Real-world Object Pose Estimation Evaluation

Synthetic-to-real genaralizability is crucial for models trained solely on synthetic data, and the accuracy of pose estimation is foundational for vision-centric interactive robotic manipulation.. Therefore, we visually represent our predictions by displaying the predicted 6D pose and 3D size in the form of a tight-oriented bounding box, as in Fig. 4.

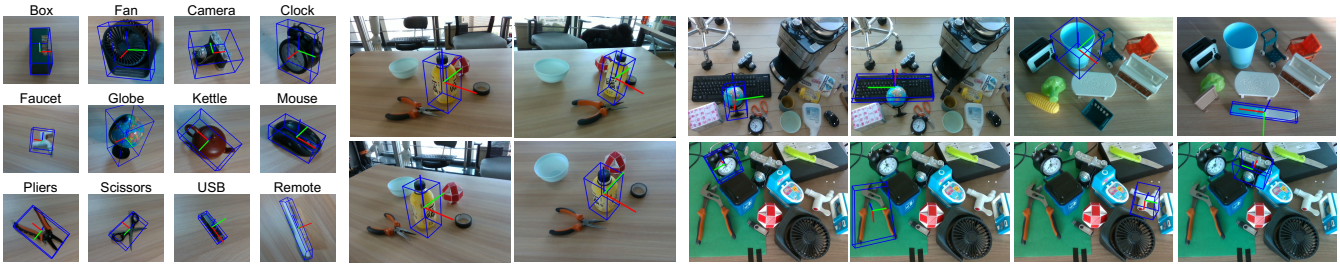
Fig. 3. **Real-world experimental objects.** We test our method using different instances from multiple tabletop-level objects, some of which are confusing in terms of color, shape, rigidity, deformability, and functionality.

To confirm the effectiveness of our model in real-world scenarios, we evaluated instances of 24 predefined tabletop-level categories within a single scene. For each instance, we captured real-world images from various viewpoints. As illustrated in Fig. 4 (a), we visually present a subset of the pose estimation results. The results are depicted using tightly oriented bounding boxes. The object is visibly positioned within and aligned with the box, showing the model's accurate estimation performance. Since the real-world test instances are not used to train the model, these results demonstrate the synthetic-to-real capability of our data rendering method and the model. Furthermore, the examples given in Fig. 4 (b) indicate that the model is capable of consistently generating accurate pose and size results, even with significant changes in viewpoint. To thoroughly evaluate the model's performance in cluttered environments with diverse backgrounds and objects, we place the target instance in a scene with numerous objects. The multi-object scene poses more challenges, as the objects are randomly placed, and some objects occlude each other. This setting allows us to assess the robustness of our method in real-world environments. As depicted in Fig. 4 (c), some objects are partially occluded, but the predicted pose and size remain accurate. These results demonstrate the effectiveness of our method and prove that using our approach allows for a fast and scalable extension of the pose estimator to multiple categories, enabling adaptability to a wider range of objects.

B. Open-ended Interactive Real-Robot Experiments

The open-ended interactive robot experiments mainly consist of two parts: instance-oriented grasping and task-oriented manipulation. Different methods and tasks share the same experimental environment and hardware.

Instance-oriented Grasping. We conducted instance-oriented grasping experiments using the Polaris framework and the following constructed baselines: 1) **RandomGrasp** randomly selects a grasping target from the instance space generated by the Polaris vision grounding module until it grasps the object requested by the user. 2) **Polaris (w/o 3DGCN)** omits the trained 3DGCN [60] used in MVPoseNet6D, where



(a) Single-object scene (b) Same object under multiple views (c) Multiple objects under the same view

Fig. 4. **Results of real-world object pose estimation.** (a) Test results of single-object scene. We present a subset of the visualization results of the pose and size estimation using the trained MVPoseNet6D model. The outcomes are represented with a tightly oriented 3D bounding box and colored XYZ-axis. (b) The scene with same object under multiple views. We show the pose of a bottle under different views. (c) The scene with multiple objects under the same view. We show the pose estimation of different objects in several cluttered scenes.

TABLE I

RESULTS OF INSTANCE-ORIENTED GRASPING

Method	Accuracy(%)	# Questions	Time(ms)	Success / Trials	Success Rate(%)
RandomGrasp	*	*	43.7	22 / 60	36.67
Ours (full model)	**93.52**	**1.36**	509.6	55 / 60	**91.67**
Ours (w/o 3DGCN)	90.07	1.41	442.7	53 / 60	88.33
Ours (w/o Light-HQSAM | w/ SAM)	87.92	1.97	849.6	49 / 60	81.67
Ours (w/o GPT-4 | w/ GPT-3.5-turbo)	93.37	3.82	552.3	55 / 60	91.67
Ours (w/o FSP | w/ FHemiSP)	82.69	1.39	497.8	42 / 60	70.00

TABLE II

RESULTS OF TASK-ORIENTED MANIPULATION

Method	Single-object Scene		Success(%)	Multi-object Scene		Success(%)	Compositional Tasks	Total Success(%)
	Pick-and-Place	Handover		Stack	Tidy			
Ours	18 / 20	17 / 20	87.50	10 / 15	12 / 15	73.33	6 / 10	**78.75**

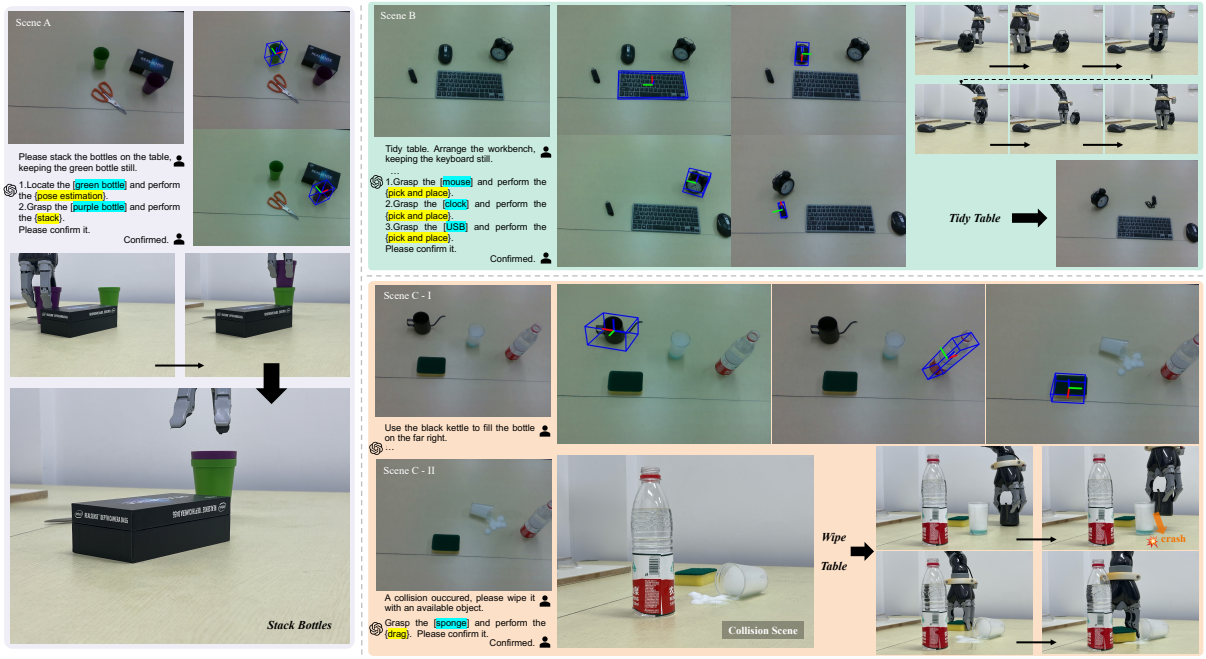


Fig. 5. **Examples of open-ended interactive real-robot experiments.** Manipulation tasks for three different base scenes are presented, including excerpts from the interaction process between the user and the LLM, the pose estimation results of the manipulated objects in different scenes, and the keyframes of the robot manipulation. **Scene A:** Stack bottles on the table. **Scene B:** Tidy the items of workbench. **Scene C:** A compositional task considering the affordance of objects after a sudden collision.

the primary function of 3DGCN is to filter out speckle and background noise of the different category-level objects point cloud captured by the depth camera. 3) **Polaris (w/o Light-HQSAM | w/ SAM)** replaces Light-HQSAM with the original pre-trained SAM [61]. 4) **Polaris (w/o GPT-4 | w/ GPT-3.5-turbo)** replaces GPT-4 with GPT-3.5-turbo [62], and

the environment's assets are provided by both the labelling model and manually by the user. The task-level prompts for the LLM(GPT-3.5-turbo) are more verbose and complex. 5) **Polaris (w/o FSP | w/ FSemiSP)** replaces Fibonacci sphere sampling with Fibonacci hemisphere sampling during synthetic data rendering. In Table I, we report the runtime, the number of questions asked, runtime, the visual accuracy (calculate the deviation between the estimated 6D pose of real-world objects and manually annotated poses), and the number of successful attempts for each method. First of all, the full model of Polaris achieves a visual accuracy of 93.52%, which demonstrates the feasibility and effectiveness of our proposed syn2real pose estimation method. The substantial increase in grasping success rate compared to RandomGrasp (from 36.67% to 91.67%) validates that the integration of our prompted LLM and vision grounding can help understand user intentions with efficient interaction states (averaging only 1.36 questions per successful grasp). The 3DGCN is trained for the classes used in our experiments to further optimize point clouds, as seen by the improvement in visual accuracy in our table scenes with plain backgrounds. The replacement of the segmentation model demonstrates that our integration of Light-HQSAM significantly reduces runtime (from 849.6ms to 509.6ms), enabling higher grounded vision accuracy at a faster operational efficiency. Furthermore, we conducted an ablation analysis on the synthetic data rendering method, where the visual accuracy obtained from hemisphere sampling decreased by about 10 percentage points. Fibonacci Sphere sampling supported a wider coverage of viewpoints, enabling the handling of more generalized scenes in real-world testing.

Task-oriented Manipulation. The high visual accuracy and grasping success rates provide a strong guarantee for task-oriented manipulation. To assess Polaris's capability in robotic manipulation, we evaluated its performance on single-object, multi-object, and compositional tasks, as shown in Table II. Quantitative results indicate that our method performs well in task-oriented manipulation, achieving success rates of 87.5% for single-object performance and 73.33% for multi-object performance, with an overall success rate of 78.75% across all tasks. Simultaneously, a qualitative analysis of the three specific examples of open-ended interactive robot manipulation in Fig. 5 reveals that our method successfully accomplished real-world tasks and effectively supported continuous interaction.

VI. Conclusion

This paper introduces Polaris, a novel framework for open-ended interactive robotic manipulation based on LLMs and syn2real visual grounding. Our syn2real pose estimation method, trained with synthetic data, performs well in real-world tests. Tabletop-level real-robot experiments provide validation of Polaris's effectiveness. We anticipate that Polaris will significantly enhance generalization across diverse, complex robotic manipulation scenarios.

References

[1] H. Lin, C. Cheang, Y. Fu, and X. Xue, "I know what you draw: Learning grasp detection conditioned on a few freehand sketches," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8417–8423.

[2] Q. Sun, H. Lin, Y. Fu, Y. Fu, and X. Xue, "Language guided robotic grasping with fine-grained instructions," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 1319–1326.

[3] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, p. eaat8414, 2019.

[4] H. Lin, Y. Fu, and X. Xue, "Pourit!: Weakly-supervised liquid perception from a single image for visual closed-loop robotic pouring," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 241–251.

[5] S. Tellex, N. Gopalan, H. Kress-Gazit, and C. Matuszek, "Robots that use language," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, pp. 25–55, 2020.

[6] C. Cheang, H. Lin, Y. Fu, and X. Xue, "Learning 6-dof object poses to grasp category-level objects by language instructions," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8476–8482.

[7] J. Huo, Q. Sun, B. Jiang, H. Lin, and Y. Fu, "Geovln: Learning geometry-enhanced visual representation with slot attention for vision-and-language navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 212–23 221.

[8] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[9] J. Mai, J. Chen, B. Li, G. Qian, M. Elhoseiny, and B. Ghanem, "Llm as a robotic brain: Unifying egocentric memory and control," *arXiv preprint arXiv:2304.09349*, 2023.

[10] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "Voxposer: Composable 3d value maps for robotic manipulation with language models," *arXiv preprint arXiv:2307.05973*, 2023.

[11] J. Wu, R. Antonova, A. Kan, M. Lepert, A. Zeng, S. Song, J. Bohg, S. Rusinkiewicz, and T. Funkhouser, "Tidybot: Personalized robot assistance with large language models," *arXiv preprint arXiv:2305.05658*, 2023.

[12] A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian, *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," in *Conference on Robot Learning*. PMLR, 2023, pp. 287–318.

[13] W. Huang, F. Xia, D. Shah, D. Driess, A. Zeng, Y. Lu, P. Florence, I. Mordatch, S. Levine, K. Hausman, *et al.*, "Grounded decoding: Guiding text generation with grounded models for robot control," *arXiv preprint arXiv:2303.00855*, 2023.

[14] N. Wake, A. Kanehira, K. Sasabuchi, J. Takamatsu, and K. Ikeuchi, "Chatgpt empowered long-step robot control in various environments: A case application," *arXiv preprint arXiv:2304.03893*, 2023.

[15] Y. Mu, Q. Zhang, M. Hu, W. Wang, M. Ding, J. Jin, B. Wang, J. Dai, Y. Qiao, and P. Luo, "Embodiedgpt: Vision-language pre-training via embodied chain of thought," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[16] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[17] J. Kim, G.-C. Kang, J. Kim, S. Shin, and B.-T. Zhang, "Gvcci: Lifelong learning of visual grounding for language-guided robotic manipulation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 952–959.

[18] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6d object pose and size estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2642–2651.

[19] N. Roy, I. Posner, T. Barfoot, P. Beaudoin, Y. Bengio, J. Bohg, O. Brock, I. Depatie, D. Fox, D. Koditschek, *et al.*, "From machine learning to robotics: challenges and opportunities for embodied intelligence," *arXiv preprint arXiv:2110.15245*, 2021.

[20] A. Romero, F. Bellas, and R. J. Duro, "A perspective on lifelong open-ended learning autonomy for robotics through cognitive architectures," *Sensors*, vol. 23, no. 3, p. 1611, 2023.

[21] S. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor, "Chatgpt for robotics: Design principles and model abilities," *2023*, 2023.

[22] J. Lin, Z. Wei, Z. Li, S. Xu, K. Jia, and Y. Li, "Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3560–3569.

[23] J. Wang, K. Chen, and Q. Dou, "Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 4807–4814.

[24] J. Lin, Z. Wei, C. Ding, and K. Jia, "Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks," in *European Conference on Computer Vision*. Springer, 2022, pp. 19–34.

[25] K. Chen and Q. Dou, "Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2773–2782.

[26] H. Lin, Z. Liu, C. Cheang, Y. Fu, G. Guo, and X. Xue, "Sar-net: shape alignment and recovery network for category-level 6d object pose and size estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6707–6717.

[27] W. Chen, X. Jia, H. J. Chang, J. Duan, L. Shen, and A. Leonardis, "Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1581–1590.

[28] Y. Di, R. Zhang, Z. Lou, F. Manhardt, X. Ji, N. Navab, and F. Tombari, "Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6781–6791.

[29] R. Zhang, Y. Di, Z. Lou, F. Manhardt, F. Tombari, and X. Ji, "Rbp-pose: Residual bounding box projection for category-level pose estimation," in *European Conference on Computer Vision*. Springer, 2022, pp. 655–672.

[30] R. Zhang, Y. Di, F. Manhardt, F. Tombari, and X. Ji, "Ssp-pose: Symmetry-aware shape prior deformation for direct category-level object pose estimation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 7452–7459.

[31] H. Xiong, Q. Li, Y.-C. Chen, H. Bharadhwaj, S. Sinha, and A. Garg, "Learning by watching: Physical imitation of manipulation skills from human videos," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 7827–7834.

[32] L. Shao, T. Migimatsu, Q. Zhang, K. Yang, and J. Bohg, "Concept2robot: Learning manipulation concepts from instructions and human demonstrations," *The International Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1419–1434, 2021.

[33] A. S. Chen, S. Nair, and C. Finn, "Learning generalizable robotic reward functions from" in-the-wild" human videos," *arXiv preprint arXiv:2103.16817*, 2021.

[34] S. Sontakke, J. Zhang, S. Arnold, K. Pertsch, E. Bıyık, D. Sadigh, C. Finn, and L. Itti, "Roboclip: One demonstration is enough to learn robot policies," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[35] Z. J. Cui, Y. Wang, N. M. M. Shafiullah, and L. Pinto, "From play to policy: Conditional behavior generation from uncurated robot data," *arXiv preprint arXiv:2210.10047*, 2022.

[36] C. Wang, L. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu, and A. Anandkumar, "Mimicplay: Long-horizon imitation learning by watching human play," *arXiv preprint arXiv:2302.12422*, 2023.

[37] J. Spisak, M. Kerzel, and S. Wermter, "Robotic imitation of human actions," *arXiv preprint arXiv:2401.08381*, 2024.

[38] P. Englert and M. Toussaint, "Learning manipulation skills from a single demonstration," *The International Journal of Robotics Research*, vol. 37, no. 1, pp. 137–154, 2018.

[39] Y. Zhu, A. Joshi, P. Stone, and Y. Zhu, "Viola: Object-centric imitation learning for vision-based robot manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 1199–1210.

[40] P. Florence, C. Lynch, A. Zeng, O. A. Ramirez, A. Wahid, L. Downs, A. Wong, J. Lee, I. Mordatch, and J. Tompson, "Implicit behavioral cloning," in *Conference on Robot Learning*. PMLR, 2022, pp. 158–168.

[41] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel, "Deep imitation learning for complex manipulation tasks from virtual reality teleoperation," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 5628–5635.

[42] S. Kumar, J. Zamora, N. Hansen, R. Jangir, and X. Wang, "Graph inverse reinforcement learning from diverse videos," in *Conference on Robot Learning*. PMLR, 2023, pp. 55–66.

[43] M. Xu, Z. Xu, C. Chi, M. Veloso, and S. Song, "Xskill: Cross embodiment skill discovery," in *Conference on Robot Learning*. PMLR, 2023, pp. 3536–3555.

[44] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

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

[46] B. Zhang and H. Soh, "Large language models as zero-shot human models for human-robot interaction," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 7961–7968.

[47] P. Wang, F. Manhardt, L. Minciullo, L. Garattoni, S. Meier, N. Navab, and B. Busam, "Demograsp: Few-shot learning for robotic grasping with human demonstration," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 5733–5740.

[48] Y.-W. Chao, Y. Xiang, *et al.*, "Fewsol: A dataset for few-shot object learning in robotic environments," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9140–9146.

[49] S. H. Kasaei, N. Shafii, L. S. Lopes, and A. M. Tomé, "Interactive open-ended object, affordance and grasp learning for robotic manipulation," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 3747–3753.

[50] Y. Yang, X. Lou, and C. Choi, "Interactive robotic grasping with attribute-guided disambiguation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8914–8920.

[51] H. Ayoobi, H. Kasaei, M. Cao, R. Verbrugge, and B. Verheij, "Explain what you see: Open-ended segmentation and recognition of occluded 3d objects," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 4960–4966.

[52] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[53] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su, "PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[54] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, L. Yi, A. X. Chang, L. J. Guibas, and H. Su, "SAPIEN: A simulated part-based interactive environment," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[55] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[56] L. Ke, M. Ye, M. Danelljan, Y. Liu, Y.-W. Tai, C.-K. Tang, and F. Yu, "Segment anything in high quality," in *NeurIPS*, 2023.

[57] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2901–2910.

[58] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 444–11 453.

[59] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *arXiv preprint arXiv:2212.08333*, 2022.

[60] Z.-H. Lin, S.-Y. Huang, and Y.-C. F. Wang, "Convolution in the cloud: Learning deformable kernels in 3d graph convolution networks for point cloud analysis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1800–1809.

[61] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," *arXiv:2304.02643*, 2023.

[62] T. Wang, Y. Li, H. Lin, X. Xue, and Y. Fu, "Wall-e: Embodied robotic waiter load lifting with large language model," *arXiv preprint arXiv:2308.15962*, 2023.