Language-Conditioned Open-Vocabulary Mobile Manipulation with Pretrained Models

Shen Tan , Dong Zhou* , Xiangyu Shao , Junqiao Wang , Guanghui Sun
Harbin Institute of Technology
shentan@stu.hit.edu.cn, dongzhou@hit.edu.cn, xiangyushao@hit.edu.cn, 23S004019@stu.hit.edu.cn,
guanghuisun@hit.edu.cn

Abstract

Open-vocabulary mobile manipulation (OVMM) that involves the handling of novel and unseen objects across different workspaces remains a significant challenge for real-world robotic applications. In this paper, we propose a novel Language-conditioned Open-Vocabulary Mobile Manipulation framework, named LOVMM, incorporating the large language model (LLM) and vision-language model (VLM) to tackle various mobile manipulation tasks in household environments. Our approach is capable of solving various OVMM tasks with free-form natural language instructions (e.g. "toss the food boxes on the office room desk to the trash bin in the corner", and "pack the bottles from the bed to the box in the guestroom"). Extensive experiments simulated in complex household environments show strong zero-shot generalization and multi-task learning abilities of LOVMM. Moreover, our approach can also generalize to multiple tabletop manipulation tasks and achieve better success rates compared to other state-of-the-art methods.

1 Introduction

As one of the key capabilities for robotic home assistance, open-vocabulary mobile manipulation (OVMM), which leverages vision cameras to navigate in the environment and execute human-like actions to manipulate unseen objects, has attracted wide attention. It is crucial for addressing real-world challenges such as object sorting and rearrangement [Zeng *et al.*, 2022], [Gan *et al.*, 2022], household cleanup [Yan *et al.*, 2021], [Wu *et al.*, 2023], and human assistance [Yenamandra *et al.*, 2023], [Stone *et al.*, 2023].

Traditionally, robotic manipulation relies on vision-based methods that use explicit, object-centric representations, including poses, categories, and instance segmentations for perception [Pan *et al.*, 2023], [Geng *et al.*, 2023a], [Xie *et al.*, 2020]. However, these approaches struggle with generalizing to unseen objects, as they often require specific training data for each scenario. Recently, end-to-end models that learn from expert demonstrations have emerged as promising alternatives [Zeng *et al.*, 2021], [Seita *et al.*, 2021], [Geng *et al.*, 2023b]. By leveraging visual observations without any explicit object information, these models are able to extract more generalizable representations across different tasks and zero-shot adapt to unseen scenarios. Yet, such methods are limited by the insufficient information provided by the single-modal data, or they may require goal images as instructions to adapt to new situations. In real-world scenarios, it is impractical to supply additional demonstrations or goal images for each new task. Thus, the model must possess the ability to open-vocabulary generalize to previously unseen tasks. An intuitive solution to this problem is grounding natural language in the manipulation policy. Natural language provides a direct interface for specifying targets and offers rich semantic information that is beneficial for more efficient learning. Although many efforts have been devoted to natural language conditioning for robotic manipulation [Kamath *et al.*, 2021], [Sharma *et al.*, 2022], these models focus on explicit representations for seen objects, while natural language instructions are mainly used for target perception, rather than helping the model learn how to manipulate in an end-to-end manner.

Recent advancements in pretrained models, especially large language models (LLMs) and vision-language models (VLMs) [Radford *et al.*, 2021], [Xue *et al.*, 2023], have demonstrated zero-shot generalization capabilities across various robotic tasks. Notably, a number of works exploit the rich semantic information that lies in different modalities by combining natural language instructions with multi-view observations [Goyal *et al.*, 2024], 3D pointclouds [Shridhar *et al.*, 2023], and action sequences [Brohan *et al.*, 2023]. These models significantly improve generalization to novel objects, but they are often restricted to single workspaces or rely on simplified environments and predefined instructions, limiting their real-world applicability.

To address these challenges, we propose natural language-conditioned open-vocabulary mobile manipulation[1] (LOVMM), a framework that integrates the LLM for reasoning and VLMs for multimodal perception, enabling both open-vocabulary navigation and end-to-end manipulation with free-form natural language instructions.

[1] The source code, dataset, and supplementary material are available at: https://github.com/shentan-shiina/LOVMM.

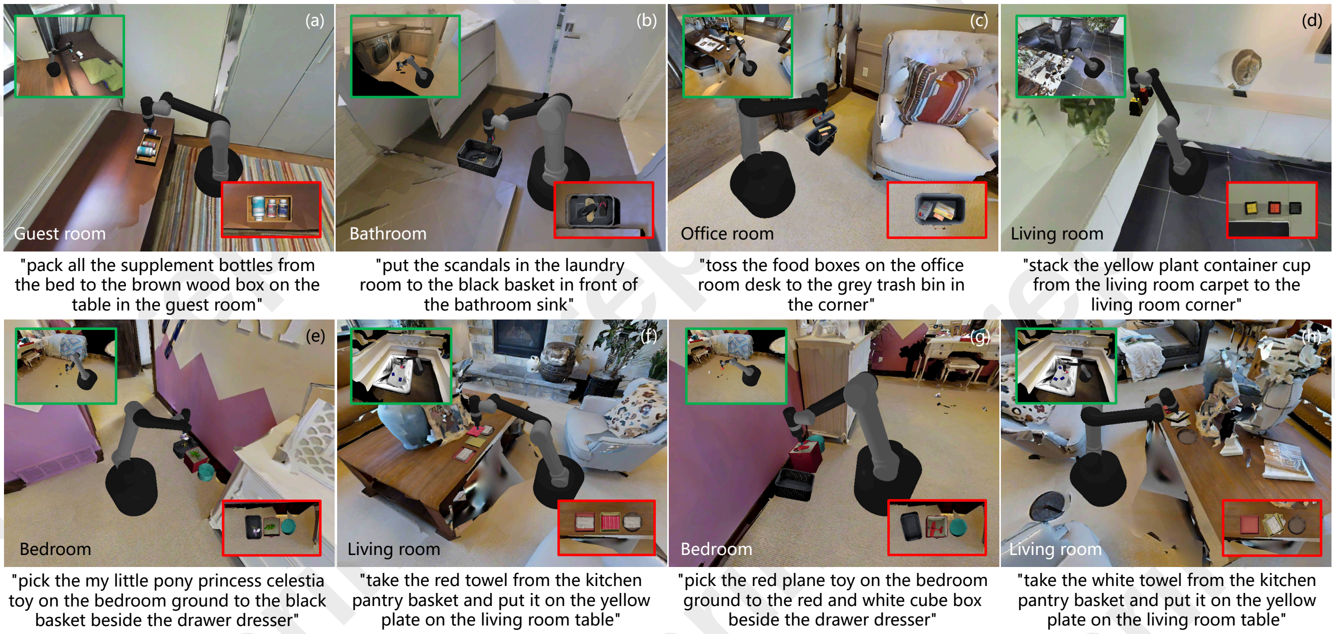
(a)
Guest room
"pack all the supplement bottles from the bed to the brown wood box on the table in the guest room"
(b)
Bathroom
"put the scandals in the laundry room to the black basket in front of the bathroom sink"
(c)
Office room
"toss the food boxes on the office room desk to the grey trash bin in the corner"
(d)
Living room
"stack the yellow plant container cup from the living room carpet to the living room corner"
(e)
Bedroom
"pick the my little pony princess celestia toy on the bedroom ground to the black basket beside the drawer dresser"
(f)
Living room
"take the red towel from the kitchen pantry basket and put it on the yellow plate on the living room table"
(g)
Bedroom
"pick the red plane toy on the bedroom ground to the red and white cube box beside the drawer dresser"
(h)
Living room
"take the white towel from the kitchen pantry basket and put it on the yellow plate on the living room table"

Figure 1: Natural language-conditioned unseen OVMM tasks. We conduct large-scale experiments based on the CLIPort benchmark in simulated indoor household scenes for 16 OVMM tasks with over 35K steps of demonstrations (see Appendix A.1 for task details).

Specifically, we employ GPT-4 [Achiam *et al.*, 2023] to parse and reason for free-form natural language instructions and VLMaps [Huang *et al.*, 2023] to construct 3D vision-language maps for navigation. Following the architecture of CLIPort [Shridhar *et al.*, 2022], we construct a two-stream model that fuses the semantic information from the vision-language representations of CLIP [Radford *et al.*, 2021] with the spatial information learned from the Transporter network [Zeng *et al.*, 2021]. The fused features from both streams are exploited to predict 6-DoF manipulation poses, facilitating efficient 3D manipulation learning and generalization to complete unseen tasks across different workspaces.

We evaluate LOVMM in simulated household environments using a mobile suction gripper robot. Our experiments are built upon the CLIPort [Shridhar *et al.*, 2022] benchmark, including over 35K steps of demonstrations across 16 different seen and unseen language-conditioned tasks, each requires open-vocabulary navigation and cross-workspace manipulation ability, as shown in Figure 1. LOVMM not only excels in multi-task learning for seen OVMM tasks, but also shows good zero-shot generalization performances for challenging unseen scenarios. In addition, the experiments further indicate that our model outperforms recent vision-based robotic manipulation methods in tabletop manipulation tasks, exhibiting more effective generalization capabilities.

The contributions of our work in this paper are summarized as follows:

- We propose a language-conditioned open-vocabulary mobile manipulation framework called LOVMM, which enables the model to handle complex OVMM tasks with free-form natural language instructions in household environments.
- We present an end-to-end 6-DoF manipulation model that exploits the joint semantic and spatial information from multimodal input for learning accurate 3D manipulation efficiently.
- A variety of experiments based on the extended benchmark of OVMM tasks are conducted. The results show that LOVMM is able to zero-shot complete diverse OVMM tasks decently and achieves superior multi-task learning and generalization performances compared to recent vision-based manipulation models.

2 Related Work

2.1 Vision-based Robotic Manipulation

Perception for vision-based robotic manipulation has traditionally relied on object-centric representations such as pose estimation [Pan *et al.*, 2023], keypoints [Liu *et al.*, 2024], and dense descriptors [Graf *et al.*, 2023]. While these methods are effective, they typically require the manipulated objects to have rich texture details or complete 3D models to extract sufficient visual features. Thus, they struggle to generalize to unseen objects due to the lack of object-specific prior knowledge.

In contrast, recent advancements in deep learning-based methods demonstrate that leveraging visual observations directly, without prior object-centric information, helps the model to better understand perception-to-action concepts and learn more generalizable manipulation policies. The Transporter network [Zeng *et al.*, 2021] finds the best object placement by cropping the image based on the sampled pick location and correlating the extracted deep visual features from both original and cropped RGB-D inputs to perform a tem-

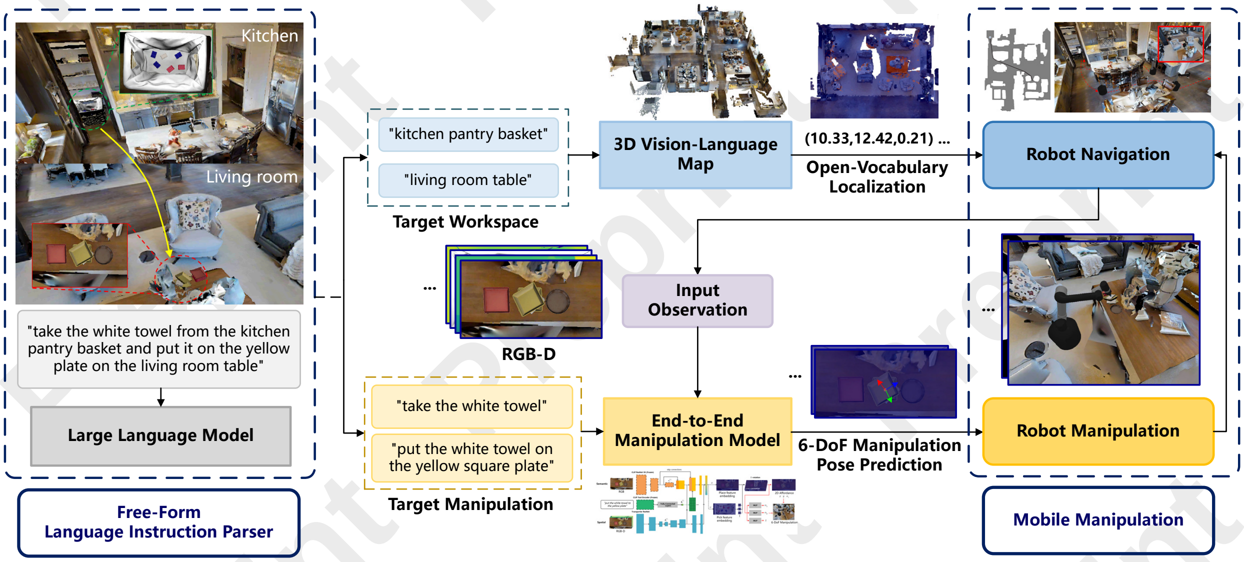
Kitchen
Living room
"take the white towel from the kitchen pantry basket and put it on the yellow plate on the living room table"
Large Language Model
Free-Form Language Instruction Parser
"kitchen pantry basket"
"living room table"
Target Workspace
RGB-D
"take the white towel"
"put the white towel on the yellow square plate"
Target Manipulation
3D Vision-Language Map
(10.33,12.42,0.21) ...
Open-Vocabulary Localization
Input Observation
End-to-End Manipulation Model
6-DoF Manipulation Pose Prediction
Robot Navigation
Robot Manipulation
Mobile Manipulation

Figure 2: Overview of LOVMM. We decompose OVMM as a series of open-vocabulary robot navigation and manipulation subtasks. Given a free-form natural language instruction, the LLM first parses the instruction to extract the target workspace and target manipulation description. The 3D vision-language map of the scene is then utilized to perform open-vocabulary localization for the robot to navigate to the specific workspace. Once the robot reaches the target location, the end-to-end manipulation model processes the target manipulation description along with RGB-D observations to predict a 6-DoF action pose for manipulation. By iteratively completing these subtasks, LOVMM enables the robot to perform complex OVMM tasks.

plate matching process. The correlation results naturally parameterize robot actions in a pick-and-place motion primitive. This formulation enables the network to learn manipulation skills efficiently but requires task-specific images to condition the policies, which limits its applicability in real-world scenarios. Further, Geng et al. [Geng *et al.*, 2023b] introduced a reinforcement learning framework that utilizes visual affordances to predict contact maps, providing a novel direction for end-to-end manipulation learning. Despite these advancements, the reliance solely on visual observations limits the perception capabilities of these methods, which makes them difficult to apply to real-world OVMM tasks. Our method, on the other hand, overcomes these limitations by integrating RGB-D images with natural language instructions as multimodal input, which can better generalize to unseen scenarios and complete diverse OVMM tasks.

2.2 Open-Vocabulary Mobile Manipulation

A number of prior works have explored how robots can solve various manipulation tasks, typically focusing on simple, single-workspace environments [Zeng *et al.*, 2021], [Seita *et al.*, 2021], [Shridhar *et al.*, 2022]. However, real-world robotic tasks often require complex mobile manipulation across multiple workspaces that involve both navigation and unseen object manipulation. For instance, a robot might need to retrieve an unfamiliar object from the kitchen and place it on the living room table, which is beyond the scope of traditional approaches. Such tasks are defined as OVMM and it remains an open problem [Yenamandra *et al.*, 2023]. Recent work [Qiu *et al.*, 2024] proposed a two-stage framework based on 3D semantic mapping and pretrained models to decompose OVMM as a series of object fetching tasks, which achieves a decent success rate in a variety of real-world tasks. Moreover, the paper [Stone *et al.*, 2023] focused on using various input modalities with VLM to solve open-world object manipulation and combined CoW [Gadre *et al.*, 2023] to address open-vocabulary navigation and manipulation. Although these works provide innovative approaches for solving OVMM tasks, they struggle to generalize to different unseen environments and are restricted to simplified task setups that only work with seen objects or a single workspace. Our approach, LOVMM, advances this field by enabling zero-shot handling of diverse OVMM tasks that involve a wide range of complex, unseen environments with novel objects across different workspaces.

2.3 Pretrained Models for Robotics

The advent of large pretrained models has sparked significant interest in applying their generalization capabilities to robotic tasks, including manipulation [Zeng *et al.*, 2021], navigation [Geng *et al.*, 2023b], and even human assistance [Kedia *et al.*, 2024]. Leveraging the strong reasoning and abstraction abilities of LLM, a number of researchers have introduced innovative approaches for grounding natural language and other different modalities into robotic learning. By decomposing high-level tasks into pretrained low-level skills with LLMs and using corresponding value functions to provide environment-specific knowledge, SayCan [Ahn *et al.*, 2022] enables real-world long-horizon robotic task planning with natural language instructions.

In parallel, vision-language models (VLMs) that enable zero-shot capabilities by training on image-text pairs have also demonstrated impressive generalization performances in various tasks [Liu *et al.*, 2023], [Kirillov *et al.*, 2023], [Yang *et al.*, 2024]. By combining pretrained VLMs with imitation learning, some works have expanded traditional language-conditioned robotic manipulation paradigm [Shridhar *et al.*, 2023], [Goyal *et al.*, 2023], [Goyal *et al.*, 2024]. Previous work [Shridhar *et al.*, 2022] proposed a language-conditioned imitation learning agent that takes advantage of both CLIP [Radford *et al.*, 2021] and Transporter [Zeng *et al.*, 2021] to learn general semantic concepts and precise spatial placement in few-shot settings. Nevertheless, it is constrained to 3-DoF manipulation tasks and it only focuses on manipulating in a fixed workspace with simple environment setups. Moreover, recent works that leverage vision-language-action models (VLAs) to directly learn generalizable robot actions are capable of tackling various manipulation tasks. However, these models often incorporate complex structures and require costly training [Kim *et al.*, 2024], [Team *et al.*, 2024]. To this end, we propose a novel framework that exploits the strong language reasoning capabilities of GPT-4 for parsing free-form natural language instructions and the rich semantic information of pretrained VLMs for open-vocabulary navigation and manipulation. This enables efficient learning of 6-DoF manipulation skills and generalization of a wide range of OVMM tasks in complex, unseen environments.

3 Proposed Method

In this section, we formulate the OVMM problem with free-form natural language instruction as input and describe LOVMM in detail. The overview of LOVMM is given in Figure 2.

3.1 Problem Formulation

The task of OVMM involves the mobile robot with vision cameras and the current environment. We formulate the problem as completing a series of subtasks as follows:

Language Instruction Parsing. Given a free-form natural language instruction $\mathbf{L}_t$ at each timestep t, it describes the target workspace $\mathbf{l}_{w_t}$ and target manipulation $\mathbf{l}_{m_t}$. We can parse the language instruction into such two parts $\mathbf{L}_t \to (\mathbf{l}_{w_t}, \mathbf{l}_{m_t})$, where each element is in text form. For example, an input instruction "toss the food boxes on the office room desk" can be parsed into the target workspace "the office room desk" and the target manipulation "toss the food boxes".

Navigating to the Target Workspace. The robot navigate to the position $\mathbf{p}_t$ of the current target workspace based on $\mathbf{l}_{w_t}$ and captures a visual observation $\mathbf{o}_t$ of the environment.

Manipulating in the Target Workspace. After reaching the target position, the problem can be considered as solving a tabletop pick-and-place manipulation subtask in the current workspace, which can be formulated as executing a manipulation policy π that outputs robot actions $\mathbf{a}_t$:

$$\pi(\mathbf{o}_t, \mathbf{l}_{m_t}) \to \mathbf{a}_t = (\mathcal{T}_{\text{pick}}, \mathcal{T}_{\text{place}}) \in \mathcal{A} \tag{1}$$

where $\mathcal{T}_{\text{pick}}$ and $\mathcal{T}_{\text{place}}$ are the poses of the robot end-effector for picking and placing actions, respectively. Both poses are defined in SE(3) for 3D manipulation. The observation $\mathbf{o}_t$ is a top-down orthographic RGB-D projection of the workspace. The target manipulation description $\mathbf{l}_{m_t}$ specifies current manipulation task. When the current manipulation subtask is completed, the robot continues to navigate to the next given target workspace and manipulate again.

By repeating to finish the sequence of navigation and manipulation subtasks, the robot naturally solves the OVMM task. In practice, we use input-action pairs $\zeta_i = \{(\mathbf{o}_1, \mathbf{l}_{m_1}, \mathbf{a}_1), (\mathbf{o}_2, \mathbf{l}_{m_2}, \mathbf{a}_2), \ldots\}$ to define each discrete-time tabletop manipulation instance, and the expert demonstration for each OVMM task can be presented as $\mathcal{D}_i = \{(\mathbf{L}_1, \mathbf{l}_{w_1}, \mathbf{p}_1, \zeta_1), (\mathbf{L}_2, \mathbf{l}_{w_2}, \mathbf{p}_2, \zeta_2), \ldots\}$.

3.2 Open-Vocabulary Navigation with Free-Form Natural Language Instruction

Given a free-form natural language instruction that specifies the OVMM task, LOVMM first uses GPT-4 as the free-form language instruction parser to interpret the target workspace $\mathbf{l}_{w_t}$ and target manipulation $\mathbf{l}_{m_t}$. Then, we follow the previous work [Huang *et al.*, 2023] and leverage the 3D reconstruction of current scene to construct a vision-language feature map matrix $Q \in \mathbb{R}^{\bar{H}\bar{W} \times C}$ using LSeg [Li *et al.*, 2022] pixel embeddings, where $\bar{H}$ and $\bar{W}$ are the size of the predefined top-down grid map, C is the length of the embedding vector. Each row of Q represents the embedding of a pixel in the map. The parsed target workspace list $\mathbf{l}_{w_t}$ is encoded with the CLIP text encoder and organized into an embedding matrix $E \in \mathbb{R}^{M \times C}$, where M represents the number of the target category. In this way, we can calculate the similarity between the given target workspace texts and the map pixels, and the highest ones indicate the most likely the pixels belong to the corresponding categories, which is formulated as:

$$\mathbf{M}_c = \text{argmax} Q \cdot E^T \tag{2}$$

where $\mathbf{M}_c \in \mathbb{R}^{\bar{H}\bar{W}}$, each element represents the label index of the target workspace categories. By choosing the most related pixels and reprojecting them to the original 3D scene map, we can localize each target workspace and obtain its position $\mathbf{p}_t$ for robot navigation.

3.3 Natural Language-Conditioned End-to-End Manipulation

Learning 6-DoF Manipulation

After reaching the target workspace, the robot is ready to perform tabletop manipulation. We first construct our model with a similar template-matching approach based on the Transporter network [Zeng *et al.*, 2021] to learn 2D planar manipulation, where $\mathcal{T}_{\text{pick}}, \mathcal{T}_{\text{place}} \in$ SE(2). Considering the above-mentioned pick-and-place policy π, we use fully convolutional networks (FCN) to model two action-value functions $\mathcal{Q}_{\text{pick}}$ and $\mathcal{Q}_{\text{place}}$. The first FCN f_{pick} takes in $\gamma_t = (\mathbf{o}_t, \mathbf{l}_{m_t})$ and outputs the pick action-value prediction $\mathcal{Q}_{\text{pick}} \in \mathbb{R}^{H \times W}$ that is used to calculate the pick action $\mathcal{T}_{\text{pick}}$:

$$\mathcal{T}_{\text{pick}} = \underset{(u,v)}{\text{argmax}}\ \mathcal{Q}_{\text{pick}}((u,v)\,|\gamma_t\,) \tag{3}$$

where (u, v) is the pixel location of the visual observation that can be mapped to the scene as a planar translation (x, y) using camera-to-robot calibration. $\mathcal{T}_{\text{pick}} \sim (u, v) \in \mathbf{o}_t$ is the pick action at the corresponding location. The second and the third FCN ψ and ϕ take in the same input γ_t and

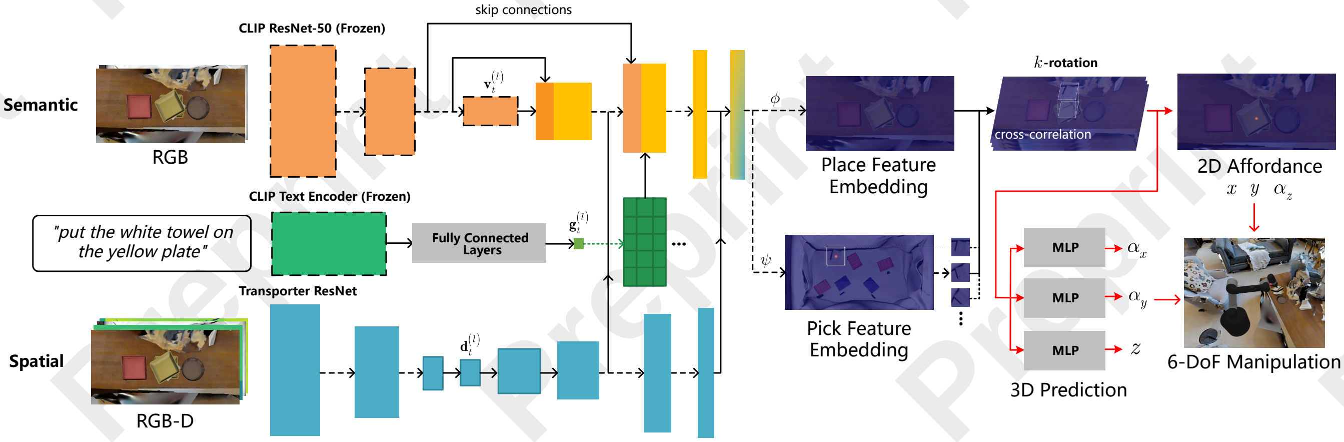

Figure 3: Architecture of the proposed end-to-end manipulation model, which adopts a two-stream architecture to fuse the visual observations and natural language instruction of the pick-and-place manipulation tasks. The fused feature embeddings are then cropped and cross-correlated to produce a 2D affordance and are further exploited with multi-layer perceptrons to predict the 6-DoF manipulation pose.

outputs two feature embeddings of shape $\mathbb{R}^{H\times W\times d}$. Then, we crop the feature embedding from ψ centered at $\mathcal{T}_{\text{pick}}$ as the query feature template to cross-correlate with the output key feature from ϕ to compute the place action-values $\mathcal{Q}_{\text{place}}$ and the corresponding place action $\mathcal{T}_{\text{place}}$:

$$\mathcal{Q}_{\text{place}}(\Delta\tau\,|\gamma_t\,,\mathcal{T}_{\text{pick}}) = ((\psi(\gamma_t)[\mathcal{T}_{\text{pick}}]) * \phi(\gamma_t))[\Delta\tau] \tag{4}$$

$$\mathcal{T}_{\text{place}} = \text{argmax}_{\Delta\tau}\,\mathcal{Q}_{\text{place}}(\,\Delta\tau|\,\gamma_t, \mathcal{T}_{\text{pick}}) \tag{5}$$

where $\psi(\gamma_t)[\mathcal{T}_{\text{pick}}]$ is the $c \times c$ partial crop of the feature embedding from ψ. Different from the original Transporter [Zeng *et al.*, 2021], we crop the feature embedding $\psi(\gamma_t)$ instead of cropping input observation $\mathbf{o}_t$ directly for a better receptive field. $\Delta\tau \in \mathrm{SE}(2)$ represents the potential planar place pose, which is discretized into k angles for the yaw rotation α_z. In this work, we use $c = 64$, $k = 36$ and $d = 3$.

With the current 2D actions $\mathcal{T}_{\text{pick}}$ and $\mathcal{T}_{\text{place}}$, we further leverage the rich spatial information embedded in the feature representations to learn 6-DoF manipulation. We apply a 1×1 convolution after the output layers of ψ and ϕ to adjust the feature channel dimension to d'. Then, by splitting the feature channels into three subsets, we utilize a separate cross-correlation and an MLP network $f(\cdot)$ for each subset to learn precise values for the remaining degrees-of-freedom, which can be formulated as follows:

$$\alpha = f(((\psi'(\gamma_t)[\mathcal{T}_{\text{pick}}]) * \phi'(\gamma_t))[\mathcal{T}_{\text{place}}]) \tag{6}$$

where α represents each remaining degree-of-freedom: the roll angle α_x, pitch angle α_y, and height z. ψ' and ϕ' are identical to ψ and ϕ except for the additional 1×1 convolution, and we use $d' = 24$. In this way, we can predict an accurate 6-DoF action pose for the manipulation target.

Two-stream Architecture

Specifically, all the FCNs f_{pick}, ψ and ϕ are constructed based on a two-stream architecture [Shridhar *et al.*, 2022] to allow for natural language conditioning and joint semantic and spatial understanding, as shown in Figure 3. The spatial information stream uses an hourglass encoder-decoder model based on the Transporter ResNet [Zeng *et al.*, 2021] network but with additional bottleneck layers for better spatial understanding, which takes in RGB-D visual observation $\mathbf{o}_t$ and outputs the feature embedding $\mathbf{d}_t^{(l)}$ at layer l. The semantic information stream leverages the pretrained CLIP ResNet-50 [Radford *et al.*, 2021] image encoder to encode the RGB input $\tilde{\mathbf{o}}_t \rightarrow \mathbf{v}_t^{(0)} : \mathbb{R}^{7\times 7\times 2048}$, and uses skip-connected upsampling decoding layers to output feature tensors $\mathbf{v}_t^{(l-1)} \rightarrow \mathbf{v}_t^{(l)} : \mathbb{R}^{h\times w\times C}$. To exploit the natural language instruction, the CLIP Transformer-based text encoder [Radford *et al.*, 2021] is utilized to get a language embedding $\mathbf{l}_{m_t} \rightarrow \mathbf{g}_t : \mathbb{R}^{1024}$. The language embedding is then downsampled and tiled with fully connected layers to produce $\mathbf{g}_t \rightarrow \mathbf{g}_t^{(l)} : \mathbb{R}^{h\times w\times C}$ such that the decoder feature embeddings of the semantic information stream can be conditioned through an element-wise product $\mathbf{v}_t^{(l)} \odot \mathbf{g}_t^{(l)}$. Then, the spatial and semantic information streams are fused with lateral connections that concatenate two feature tensors, and a 1×1 convolution is applied to adjust the channel dimension of the output feature embedding, which is formulated as:

$$[\mathbf{v}_t^{(l)} \odot \mathbf{g}_t^{(l)}; \mathbf{d}_t^{(l)}] : \mathbb{R}^{h\times w\times C_{\mathbf{v}}+C_{\mathbf{d}}} \rightarrow \mathbb{R}^{h\times w\times C_{\mathbf{v}}} \tag{7}$$

where $C_{\mathbf{v}}$ and $C_{\mathbf{d}}$ are the channel sizes of the semantic and spatial tensors.

Finally, we train the end-to-end manipulation model through imitation learning from expert demonstrations $\mathcal{D}_i$. We first randomly sample an input-action pair ζ_i and then supervise the model with one-hot pixel encodings of the expert actions $Y_{\text{pick}} : \mathbb{R}^{H\times W\times k}$ and $Y_{\text{place}} : \mathbb{R}^{H\times W\times k}$ in an end-to-end manner. The model is trained with cross-entropy loss for 2D manipulation, which is defined as follows:

$$\mathcal{L}_{\text{2D}} = -\mathbb{E}_{Y_{\text{pick}}}[\log\mathcal{V}_{\text{pick}}] - \mathbb{E}_{Y_{\text{place}}}[\log\mathcal{V}_{\text{place}}] \tag{8}$$

where $\mathcal{V}_{\text{pick}} = \text{softmax}(\mathcal{Q}_{\text{pick}}((u, v)|\,\gamma_t))$ and $\mathcal{V}_{\text{place}} = \text{softmax}(\mathcal{Q}_{\text{place}}(\Delta\tau|\,\gamma_t, \mathcal{T}_{\text{pick}}))$. For each remaining degree-of-freedom prediction for 3D manipulation, we use a Huber

Method	Task-A			Task-B			Task-C			Task-D		
	1	10	100	1	10	100	1	10	100	1	10	100
LOVMM	52.7	19.7	**53.5**	**91.3**	80.6	83.9	68.0	**73.1**	64.6	5.6	1.7	**6.3**
Method	Task-E			Task-F			Task-G			Task-H		
	1	10	100	1	10	100	1	10	100	1	10	100
LOVMM	63.0	34.0	**69.0**	43.1	**82.9**	82.5	59.0	34.0	**62.0**	44.3	**72.7**	63.9

Table 1: Seen OVMM tasks evaluation results.

loss to train the MLP, which is formulated as follows:

$$\mathcal{L}_{\text{3D}} = \begin{cases} 0.5(\hat{\theta} - \theta)^2, & \text{if } |\hat{\theta} - \theta| < 1 \\ |\hat{\theta} - \theta| - 0.5, & \text{otherwise} \end{cases} \tag{9}$$

where $\hat{\theta}$ is the predicted value of each remaining degree-of-freedom, θ is the true value.

4 Experiment

In this section, we evaluate the performance of LOVMM by conducting extensive simulated experiments in household environments. In addition, we further explore the effectiveness of our model by comparing it against baseline methods in natural language-conditioned open-vocabulary tabletop manipulation tasks. All models are trained on 4 NVIDIA RTX 4090 GPUs.

4.1 LOVMM Performance for OVMM Tasks

We design 16 different natural language-conditioned OVMM tasks in various indoor scenes with household environments. The tasks are divided into 8 simpler seen tasks for training, and 8 more challenging unseen tasks for testing, as illustrated in Figure 1. See Appendix A.1 for details on the task setup. The models are trained for 600K steps across all seen tasks using $n = 1, 10, 100$ expert demonstrations in multi-task settings following CLIPort benchmark [Shridhar *et al.*, 2022]. Then we evaluate the models on 100 seen tasks and use the best validation model to test on 100 unseen tasks. The task success rate (TSR) is adopted to assess the model performance.

Performance for Seen OVMM Tasks

The constructed OVMM seen tasks are shown in Appendix A.1 and the detailed evaluation results of all seen OVMM tasks are presented in Table 1. It is obvious that LOVMM can decently solve most of the tasks, with the best 91.3% TSR for *Task-B*. Specifically, LOVMM trained with 100 task demonstrations outperforms other models in half of the tasks, with a performance of 53.5% for *Task-A*, which is over 30.0% higher than the model trained with 10 demonstrations. We can also calculate the highest 60.7% average TSR across all seen tasks of LOVMM using 100 demonstrations, showcasing its efficient multi-task learning and accurate manipulation abilities. Furthermore, when trained with limited demonstrations, LOVMM still shows decent performances, reaching 91.3% and 59.0% TSRs for *Task-B* and *Task-G*, respectively. LOVMM is able to reach a 53.4% average TSR using only 1 seen task demonstration, which further demonstrates the efficient learning ability of LOVMM to grasp diverse manipulation skills.

However, a notable performance drop can be observed in some of the evaluated tasks when LOVMM is trained with 10 demonstrations. We hypothesize such a model behavior is caused by the imbalanced dataset and the random sampling strategy. See Appendix A.5 for further discussion.

Method	Task-I	Task-J	Task-K	Task-L
LOVMM	7.3	21.2	21.0	3.9
Method	Task-M	Task-N	Task-O	Task-P
LOVMM	8.9	7.8	9.1	3.2

Table 2: Unseen OVMM tasks evaluation results.

Performance for Unseen OVMM Tasks

Based on the evaluation performances for seen OVMM tasks, we choose LOVMM trained with 100 expert demonstrations to test on unseen OVMM tasks. The evaluation results are presented in Table 2. See Appendix A.2 for detailed evaluation results. Compared with seen tasks, the performances of LOVMM decrease as the unseen tasks are inherently more difficult and involved with diverse and challenging environments. It shows that our models can still generalize to complete many of them, reaching over 20.0% TSR for tasks that require strong open-vocabulary manipulation capabilities, such as *Task-J* and *Task-K*. Notably, even though *Task-I* requires generalization of unseen object categories and precise manipulation to solve, our model achieves the best performance of 7.3% TSR, which shows the zero-shot open-vocabulary generalization and cross-workspace manipulation capabilities of LOVMM. Our model also achieves nearly 10.0% performance when generalizing to *Task-O*, showcasing its 6-DoF manipulation learning ability. Even for extremely challenging tasks like *Task-M*, which not only needs precise language parsing for open-vocabulary localization but also requires correct understanding of the corresponding pick-and-place targets, LOVMM manages to complete some instances. Covering all tasks, LOVMM is able to reach a 10.2% average TSR using 100 task demonstration, which demonstrates that LOVMM has strong multi-task learning ca-

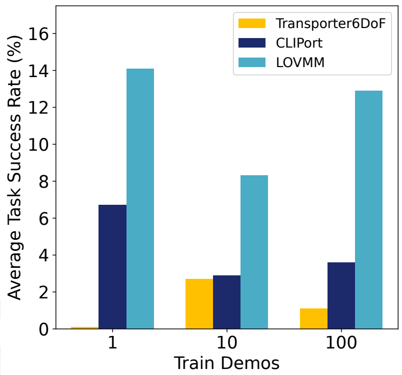

Figure 4: Average task success rates for tabletop manipulation tasks.

pabilities to leverage limited semantic information and manipulation concepts across different tasks efficiently to zero-shot generalize to unseen environments and novel object attributes.

In general, LOVMM is capable of learning manipulation skills efficiently and zero-shot generalizing to complete mobile manipulation in diverse, complex environments, providing a feasible solution for addressing a wide range of language-conditioned OVMM tasks with a one-for-all multi-task model.

4.2 Performance Comparison for Tabletop Manipulation Tasks

To further validate the manipulation performance of our model, we compare our model against the image-conditioned Transporter with 6-DoF placing [Zeng *et al.*, 2021] and language-conditioned CLIPort [Shridhar *et al.*, 2022] on 100 unseen tabletop manipulation tasks under the same multi-task training settings. The summarized average task success rates are shown in Figure 4. See Appendix A.3 for detailed evaluation results.

It is obvious in Figure 4 that our multi-task model performs exceptionally better than all the other baselines. To be specific, while Transporter6DoF can hardly reach only 0.1% using 1 task demonstration, LOVMM reaches nearly 15.0% TSR, surpassing the second-best CLIPort by more than double. As the number of training demonstrations increases, the average TSR of LOVMM first decreases to about 8.0% but then recovers to over 12.0% when trained with 100 demonstrations, which is 10 times better than the performance of Transporter6DoF. In contrast, the performances of CLIPort and Transporter6DoF show limited improvement as the number of expert demonstrations increases, with TSRs lower than 5.0%. These results further demonstrate that our LOVMM model has superior capabilities for efficient multi-task learning and adapting to novel unseen tasks in complex environments.

4.3 Ablation Study

To evaluate the impact of various components of the proposed end-to-end manipulation model, we conduct a series of ablation studies. Specifically, we examine the effect of (i) removing data augmentation (details of data augmentation are presented in Appendix A.4), (ii) cropping the input observation instead of the feature embedding, (iii) removing additional bottleneck layers, and (iv) using only 3-DoF manipulation. All ablation models are trained from scratch using 100 demonstrations and evaluated on 100 seen tasks. The results are summarized in Table 3. It clearly shows that each component contributes to the performance of LOVMM. Removing data augmentation results in a substantial 19.1% drop in average TSR, highlighting its crucial role in improving the model's learning efficiency and generalization. Moreover, cropping the observation directly may lead to the loss of receptive field, which makes the network less capable of perceiving complex environments, thus causing a performance decrease. Similarly, the additional bottleneck layers help the model extract richer perceptual information from the observations to grasp manipulation skills. The remaining degrees-of-freedom are also essential for the model to solve tasks that require 6-DoF placements such as *Task-P*.

Method	Average TSR
W/o data augmentation	34.3
Crop for input observation	50.3
No additional bottleneck layers	48.6
3-DoF manipulation	47.9
Original	**53.4**

Table 3: Ablation Study on LOVMM for OVMM Tasks.

Overall, LOVMM demonstrates the ability to efficiently learn multi-task manipulation policies and zero-shot generalize to finish challenging OVMM tasks in unseen scenarios. The ablation results also highlight the importance of the proposed components in our end-to-end manipulation model, which are essential for LOVMM to achieve an excellent performance.

5 Conclusion

In this paper, we formulate the OVMM task as completing a series of navigation and tabletop manipulation subtasks and propose a novel pretrained model-based natural language-conditioned open-vocabulary mobile manipulation framework, LOVMM, incorporating an end-to-end manipulation model for efficient multi-task manipulation learning. Our framework enables free-form natural language instruction input and can learn generalizable policies to tackle diverse OVMM tasks that involve novel, unseen environments and objects. Extensive experiments simulated in household environments show the advancement of LOVMM, and it achieves an overall better zero-shot generalization performance at solving tabletop manipulation tasks compared to other recent manipulation models.

Acknowledgements

This study was supported by the National Natural Science Foundation of China through grant No. 62403162 and the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government via the Hong Kong Centre for Logistics Robotics.

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Ahn *et al.*, 2022] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[Brohan *et al.*, 2023] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[Gadre *et al.*, 2023] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23171–23181, 2023.

[Gan *et al.*, 2022] Chuang Gan, Siyuan Zhou, Jeremy Schwartz, Seth Alter, Abhishek Bhandwaldar, Dan Gutfreund, Daniel LK Yamins, James J. DiCarlo, Josh McDermott, and Antonio Torralba. The threedworld transport challenge: A visually guided task-and-motion planning benchmark towards physically realistic embodied ai. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8847–8854. IEEE, 2022.

[Geng *et al.*, 2023a] Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7081–7091, 2023.

[Geng *et al.*, 2023b] Yiran Geng, Boshi An, Haoran Geng, Yuanpei Chen, Yaodong Yang, and Hao Dong. Rlafford: End-to-end affordance learning for robotic manipulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5880–5886. IEEE, 2023.

[Goyal *et al.*, 2023] Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. Rvt: Robotic view transformer for 3d object manipulation. In *Conference on Robot Learning*, pages 694–710. PMLR, 2023.

[Goyal *et al.*, 2024] Ankit Goyal, Valts Blukis, Jie Xu, Yijie Guo, Yu-Wei Chao, and Dieter Fox. Rvt-2: Learning precise manipulation from few demonstrations. *arXiv preprint arXiv:2406.08545*, 2024.

[Graf *et al.*, 2023] Christian Graf, David B. Adrian, Joshua Weil, Miroslav Gabriel, Philipp Schillinger, Markus Spies, Heiko Neumann, and Andras Gabor Kupcsik. Learning dense visual descriptors using image augmentations for robot manipulation tasks. In *Conference on Robot Learning*, pages 871–880. PMLR, 2023.

[Huang *et al.*, 2023] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual Language Maps for Robot Navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10608–10615, May 2023.

[Kamath *et al.*, 2021] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

[Kedia *et al.*, 2024] Kushal Kedia, Atiksh Bhardwaj, Prithwish Dan, and Sanjiban Choudhury. Interact: Transformer models for human intent prediction conditioned on robot actions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 621–628. IEEE, 2024.

[Kim *et al.*, 2024] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, and Wan-Yen Lo. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[Li *et al.*, 2022] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022.

[Liu *et al.*, 2023] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[Liu *et al.*, 2024] Fangchen Liu, Kuan Fang, Pieter Abbeel, and Sergey Levine. Moka: Open-vocabulary robotic manipulation through mark-based visual prompting. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

[Pan *et al.*, 2023] Chuer Pan, Brian Okorn, Harry Zhang, Ben Eisner, and David Held. Tax-pose: Task-specific cross-pose estimation for robot manipulation. In *Conference on Robot Learning*, pages 1783–1792. PMLR, 2023.

[Qiu *et al.*, 2024] Dicong Qiu, Wenzong Ma, Zhenfu Pan, Hui Xiong, and Junwei Liang. Open-vocabulary mobile manipulation in unseen dynamic environments with 3d semantic maps. *arXiv preprint arXiv:2406.18115*, 2024.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[Seita *et al.*, 2021] Daniel Seita, Pete Florence, Jonathan Tompson, Erwin Coumans, Vikas Sindhwani, Ken Goldberg, and Andy Zeng. Learning to rearrange deformable cables, fabrics, and bags with goal-conditioned transporter networks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4568–4575. IEEE, 2021.

[Sharma *et al.*, 2022] Pratyusha Sharma, Balakumar Sundaralingam, Valts Blukis, Chris Paxton, Tucker Hermans, Antonio Torralba, Jacob Andreas, and Dieter Fox. Correcting robot plans with natural language feedback. *arXiv preprint arXiv:2204.05186*, 2022.

[Shridhar *et al.*, 2022] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, pages 894–906. PMLR, 2022.

[Shridhar *et al.*, 2023] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023.

[Stone *et al.*, 2023] Austin Stone, Ted Xiao, Yao Lu, Keerthana Gopalakrishnan, Kuang-Huei Lee, Quan Vuong, Paul Wohlhart, Sean Kirmani, Brianna Zitkovich, Fei Xia, et al. Open-world object manipulation using pre-trained vision-language models. *arXiv preprint arXiv:2303.00905*, 2023.

[Team *et al.*, 2024] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

[Wu *et al.*, 2023] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. TidyBot: Personalized robot assistance with large language models. *Autonomous Robots*, 47(8):1087–1102, December 2023.

[Xie *et al.*, 2020] Christopher Xie, Yu Xiang, Arsalan Mousavian, and Dieter Fox. The best of both modes: Separately leveraging rgb and depth for unseen object instance segmentation. In *Conference on Robot Learning*, pages 1369–1378. PMLR, 2020.

[Xue *et al.*, 2023] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. ULIP: Learning a unified representation of language, images, and point clouds for 3D understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1179–1189, 2023.

[Yan *et al.*, 2021] Zhi Yan, Nathan Crombez, Jocelyn Buisson, Yassine Ruichck, Tomas Krajnik, and Li Sun. A quantifiable stratification strategy for tidy-up in service robotics. In *2021 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*, pages 182–187. IEEE, 2021.

[Yang *et al.*, 2024] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024.

[Yenamandra *et al.*, 2023] Sriram Yenamandra, Arun Ramachandran, Karmesh Yadav, Austin S. Wang, Mukul Khanna, Theophile 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, editors, *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 1975–2011. PMLR, 06–09 Nov 2023.

[Zeng *et al.*, 2021] Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, and Vikas Sindhwani. Transporter networks: Rearranging the visual world for robotic manipulation. In *Conference on Robot Learning*, pages 726–747. PMLR, 2021.

[Zeng *et al.*, 2022] Andy Zeng, Shuran Song, Kuan-Ting Yu, Elliott Donlon, Francois R. Hogan, Maria Bauza, Daolin Ma, Orion Taylor, Melody Liu, Eudald Romo, Nima Fazeli, Ferran Alet, Nikhil Chavan Dafle, Rachel Holladay, Isabella Morona, Prem Qu Nair, Druck Green, Ian Taylor, Weber Liu, Thomas Funkhouser, and Alberto Rodriguez. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. *The International Journal of Robotics Research*, 41(7):690–705, June 2022.