# Broadcasting Support Relations Recursively from Local Dynamics for Object Retrieval in Clutters

Yitong Li*[1,2] Ruihai Wu*[1,4]
Haoran Lu[1,4] Chuanruo Ning[1,4] Yan Shen[1,4] Guanqi Zhan [3] Hao Dong [1,4]
[1]CFCS, School of CS, PKU [2]Weiyang College, THU [3]University of Oxford
[4]National Key Laboratory for Multimedia Information Processing, School of CS, PKU

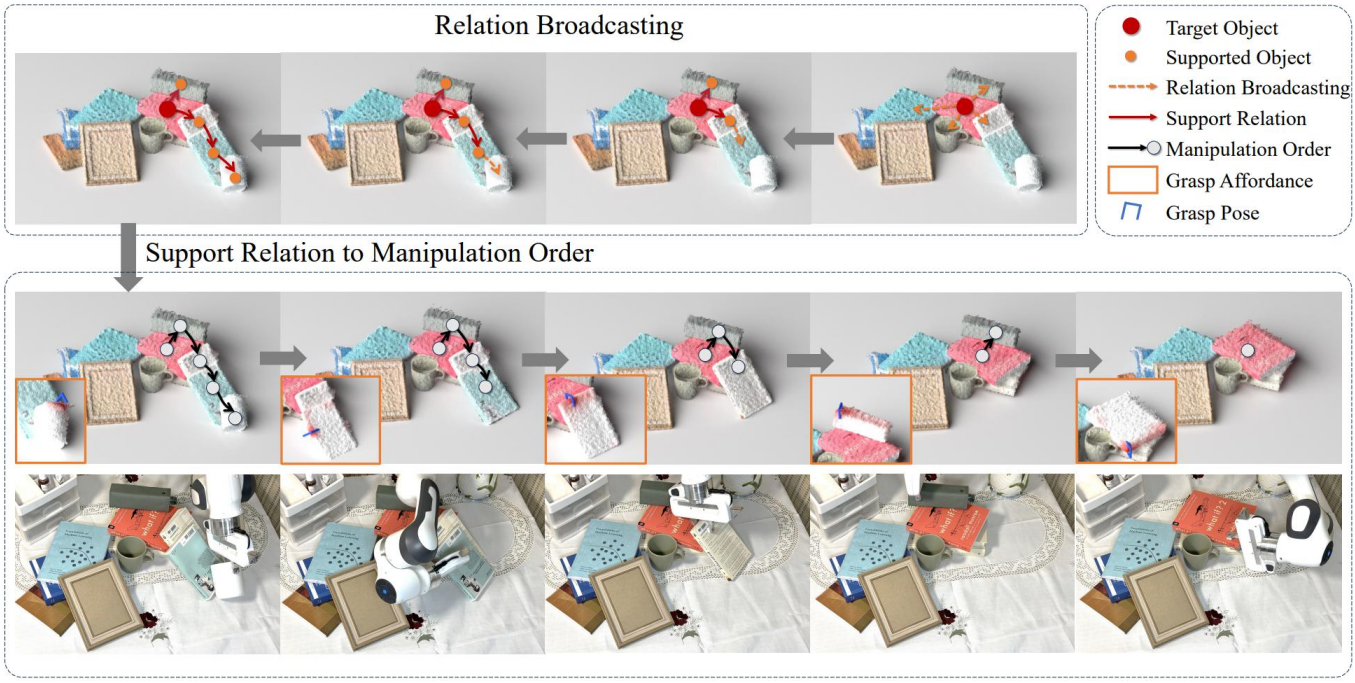


Fig. 1: **Our Proposed Framework** broadcasts the support relations recursively from the target object using local dynamics between adjacent objects, and uses the support relation graph to efficiently guide the step-by-step target object retrieval.

***Abstract*—In our daily life, cluttered objects are everywhere, from scattered stationery and books cluttering the table to bowls and plates filling the kitchen sink. Retrieving a target object from clutters is an essential while challenging skill for robots, for the difficulty of safely manipulating an object without disturbing others, which requires the robot to plan a manipulation sequence and first move away a few other objects supported by the target object step by step. However, due to the diversity of object configurations (*e.g.*, categories, geometries, locations and poses) and their combinations in clutters, it is difficult for a robot to accurately infer the support relations between objects faraway with various objects in between. In this paper, we study retrieving objects in complicated clutters via a novel method of recursively broadcasting the accurate local dynamics to build a support relation graph of the whole scene, which largely reduces the complexity of the support relation inference and improves the accuracy. Experiments in both simulation and the real world demonstrate the efficiency and effectiveness of our method.**

*Equal contribution.

## I. Introduction

Cluttered objects [57, 12, 54], such as piled books on desks and cluttered objects in kitchen, are everywhere in our daily life. To retrieve a target object in the complicated clutter [67, 14, 20, 74], *e.g.*, retrieving a book from clutters of books and stationeries, is an essential capability for future robots to assist human in various scenarios.

Compared to object-centric manipulation tasks (such as grasping a bottle, opening a drawer or closing a door), cluttered objects manipulation like retrieving is much more challenging for many reasons. One of the most important reasons is that, it requires safe manipulation, which means when manipulating the target object, other objects should not be collided. For example, to retrieve a plate in the sink filled with plates, bowls and glasses, the robot should avoid making other objects broken.

To achieve this goal, the manipulation will be long-horizon.

More specifically, the robot needs to have the planning ability to first move a few other objects away step by step, and then manipulate the target object. For the object retrieval task in the clutters, the relations that the robot should be aware of, is the supporting relations between objects. That's to say, to retrieve a target objects, other objects that are directly or indirectly supported by it should be first moved away safely.

However, due to the complexity of the clutters in terms of various combinations of diverse categories and shapes of objects in different positions and poses, it is very difficult to accurately infer support relations in the complicated scene and draw an accurate and efficient manipulation plan based on that. For example, when the two objects are distant with many objects in between, it is very difficult for a network to predict the dynamics of the other object, as the dynamics is transferred complicatedly by the chained objects in between.

To tackle this problem, we leverage the property of dynamics models that, local dynamics is much more accurate and easier to predict than the dynamics between two distant objects, and thus propose to infer to broadcast the support relations recursively from the target objects to more and more faraway objects. More specifically, we first infer the support relations between target and its adjacent object using the accurate local dynamics predictor. When some of these objects are inferred to be supported by the target object, we execute the local dynamics predictor on these objects and find new objects supported by them, which are also indirectly supported by the target object. We apply this process recursively and gradually build the support relation graph containing all the objects that should be retrieved before retrieving the target (Figure 1, first row). From this directed acylic graph (DAG), we can easily retrieve objects from those non-outdgree objects in the graph (Figure 1, second the third row).

To evaluate our framework in cluttered object retrieval, especially in complicated clutters, while previous works on supporting relation inference for robotic manipulation use environments with relatively simple object geometries [32, 17, 42, 33, 16] or in the lack of clutters complexities [40, 39, 41, 66], we propose a new evaluation environment with combinations of thousands of different objects into realistic scenarios. In this environment, extensive experiments demonstrate that our proposed framework outperforms previous works that directly infer object relations by a large margin.

In short, in this paper, we make the following contributions:

- we propose to leverage the accurate local dynamics and broadcast it recursively to study object support relations for the retrieval task in clutters;
- we propose a novel system with novel designs to efficiently build the support relation graphs that can guide downstream object retrieval task;
- extensive experiments on diverse realistic scenarios and comprehensive metrics demonstrate the superiority of our proposed framework;

## II. Related Work

### A. Support Relations Inference

Inferring support relations [51, 13, 15, 35, 42, 33] is important in object relation inference [28, 78, 60, 27] in computer vision and robotics community. A dataset [51] manually annotates support relation of different objects or regions for RGBD images. Following this, a series of works aim to infer support relations given different forms of input, including RGB [76, 82, 35, 36, 37] or RGBD [13, 69, 71, 80] images, and 3D models [15]. In contrast, robotic community formulates support relation inference problem based on both object relations and manipulation. However, object geometries [32, 17, 42, 33, 16] and clutter complexities [40, 39, 41, 66] are relatively simple. Our study infers the support relations for manipulation via the broadcasting of dynamics models, and proposes a new evaluation benchmark consisting of more diverse objects with more complicated geometry.

### B. Cluttered Objects Manipulation

In the realm of robotic manipulation, addressing the challenge of interacting with objects in cluttered scenes has garnered significant attention due to its practical implications for real-world applications ranging from object grasping [54, 57, 75, 79, 5, 30, 7, 45, 2, 10, 77], retrieval [67, 14, 20, 74, 8, 72, 56], to rearrangement [12, 6, 21, 55]. Some works leverage visual grounding [67, 30, 77] or object detection [22, 48] technique to comprehend cluttered scenes, while others advocate an end-to-end framework for direct manipulation pose prediction [2, 10, 34, 45, 54]. Among them, a subset of works [62, 57, 5, 58, 23] focuses on learning manipulation affordance as guidance for subsequent manipulation pose generation, significantly enhancing both efficiency and accuracy. However, prior works lack explicitly consideration for the relationships within cluttered scenes, and hardly impose constraints on the movement of other objects. Our study proposes the use of graphs to represent cluttered scenes and support relations between objects, which aids in safe object manipulation without disturbing others.

### C. Dynamics Models for Robotic Manipulation

Building dynamics models has been a promising approach in robot systems, which plans the manipulation through predicting the future state of objects under different actions. Finding the suitable representations for different objects is the core task in these model-based methods [70, 59, 4, 47]. Previous dynamics models made advancements by discovering appropriate abstraction for different objects, such as particle representations for deformable objects [49, 50, 24, 43, 31, 29, 64] and adjacent object interactions [4], video predictions for rigid objects [19, 11, 65, 73, 1], pixel representations for granular objects [59, 53]. However, the suitable form of dynamics models for cluttered objects are still under-explored. Our method devises the particle representation for modeling the movement of each object and uses graph representation to model the relation among cluttered objects, which effectively establishes the dynamics model for manipulation.

## III. Problem Formulation

Given a partially scanned 3D point cloud $S \in \mathbb{R}^{N\times 3}$ of a clutter containing $n$ objects $O_1, O_2, \ldots, O_n$ (where $O_j \in \mathbb{R}^{N_j\times 3}$ for each object $O_j$), with a target object $O_t$ ($t \in \{1, 2, \ldots, n\}$), the goal for the robot is to sequentially remove occluded objects that are directly or indirectly supported by the target object $O_t$, and finally retrieve $O_t$ from the clutter. Each manipulation action should be safe, meaning that the manipulation of one object should not result in displacements of other objects in their positions and poses.

Specifically, at time step $i$, the robot takes in the current scene $S_i$ and executes a manipulation action $a_i$. Each action $a_i$ safely takes out one object $O_{k_i}$, where $k_i \in \{1, 2, \ldots, n\}$ denotes the index of the manipulated object. The overall manipulation sequence involves the robot executing actions $(a_1, a_2, \ldots, a_{l-1})$ in the initial $l-1$ steps to sequentially remove $l-1$ occluded objects supported by the target object $O_t$, and the final action $a_m$ at step $l$ is to retrieve the target $O_t$, note that $O_{k_m} = O_t$. Each action $a_i$ is represented as $(p_i, r_i, d_i)$, where $p_i \in \mathbb{R}^3$ denotes the grasp point on the manipulated object $O_{k_i}$, $r_i \in SO(3)$ represents the pose for grasping at $p_i$, and $d_i \in \mathbb{R}^3$ is the direction for retrieving $O_{k_i}$ after grasping at $p_i$.

## IV. Method

### A. Motivation and Overview

As described in the **Introduction** section, directly modelling the support relations between any two objects in clutters is difficult and inaccurate, as object relations between two distant objects could be highly complicated and hard to predict because of the chained objects in between.

To tackle this problem, we build the whole support relation graph of the cluttered objects by broadcasting the more accurate local dynamics between adjacent objects recursively (Section IV-B and IV-E), with the assistance of *Retrieval Direction Predictor* (Section IV-C) and *Local Dynamics Predictor* (Section IV-D). Guided by the support relation graph, the robot can estimate the manipulation affordance (Section IV-F) and execute the retrievals step by step.

### B. General Idea for Support Graph Generation

To effectively represent complex scenes with multiple objects and their support relations, we adopt a directed acyclic graph (DAG) $\mathcal{G}$, where vertex $v_i$ represents object $O_i$ and edge $e_{ij}$ represents that object $O_i$ supports object $O_j$. Intuitively, to safely retrieve a target object $O_t$ without disturbing others, we need to estimate which objects are directly or indirectly supported by it. On the other hand, objects with no support relations to $O_t$ will minimally influence the retrieval process. Therefore, our focus lies in the hierarchical support structure centered around the target object $O_t$, and we construct a subgraph $\mathcal{G}_s$ that represents the support relations among cluttered objects, named as the *Support Graph*. This subgraph enables the derivation of a feasible retrieval sequence based on the spatial relationships outlined in the *Support Graph*.

To set up the directed acyclic subgraph $\mathcal{G}_s$, there are three important steps: *First,* we build the nodes representing the objects in the clutter, ensuring that each object can be retrieved in a way that causes as less disturbance as possible. This necessitates optimizing the retrieval direction to avoid collisions with other objects. Therefore, we propose the *Retrieval Direction Predictor* (Section IV-C), to generate and evaluate the optimal retrieval directions for each object. *Subsequently*, we build the edges between different nodes, which indicate the presence of support relation between two objects. For this purpose, we introduce the *Local Dynamics Predictor* (Section IV-D), which predicts whether a given action on an object would cause another object to lose support and displace. *Finally*, utilizing the two modules for building nodes and edges, we employ the *Clutter Solver* (Section IV-E) to recursively broadcast the inter-object relationships from the target object, thereby constructing the entire graph.

### C. Retrieval Direction Predictor

When retrieving an object in clutter, different retrieval directions, denoted as $d \in \mathbb{R}^3$ could lead to different results, *i.e.* different displacements of its adjacent objects. To minimize the displacements or collisions of adjacent objects, we introduce *Retrieval Direction Predictor* to propose the optimal retrieval direction for the object that can avoid the movements of other objects to guarantee the safety and efficiency in manipulation for the less retrieval steps.

To achieve this goal, we propose two submodules in this predictor, *Direction Proposal Module* and *Direction Scoring Module*. The *Direction Proposal Module* aims to propose direction candidates that will lead to minimal movements of other objects, and the *Direction Scoring Module* further scores the direction candidates, and select the best action direction.

In the *Direction Proposal Module*, for each object point cloud $O_i$ and its scene point cloud $S$, we use PointNet++ [44] to respectively extract their features $f_{O_i}$ and $f_S$. We sample $q$ various retrieval directions $D = \{d_1, d_2, \ldots d_q\}$ and obtaining the corresponding ground truth movement scores of other objects $\{m_1, m_2, \ldots m_q\}$ (detail described in appendix). For the retrieval directions with movement scores higher than a thresh $th_m$, we employ a conditional variational autoencoder (cVAE) [52] as the direction proposal model $F_{DP}$ to efficiently model the distribution of these good retrieval directions. Specifically, the cVAE takes the feature concatenation of $f_{O_i}$ and $f_S$ as the condition, encodes an retrieval direction $d_i$ to a latent $z_i$, and reconstructs $z_i$ into the direction $d_i'$. We employ the L1-loss between $d_i'$ and $d_i$ as the reconstruction loss:

$$L_{recon} = |d_i - F_{DP}(f_{O_i}, f_S, d_i)|. \tag{1}$$

A KL loss between $z_i$ and Gaussian distribution is applied to guarantee the sampling ability from an Gaussian noise $z$ to a direction in the promising directions distribution.

$$L_{KL} = \sum_{z_i} p(z_i) log(\frac{p(z_i)}{G(z_i)}), \tag{2}$$

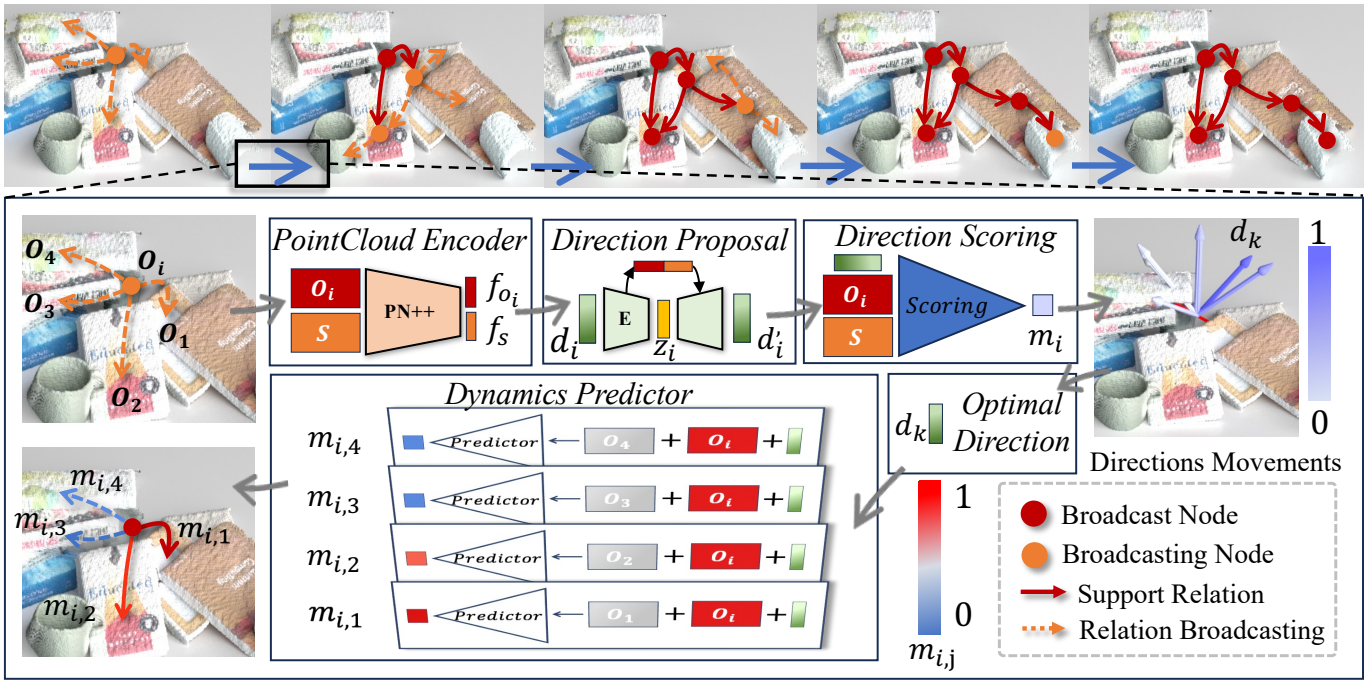


Fig. 2: **Our Proposed Framework.** The first row shows the **Recursive Broadcasting** process of support relations via local dynamics. To infer the local dynamics starting from an object $O_i$, our framework first selects the optimal retrieval direction using the *Direction Scoring Module* from the direction candidates proposed by the *Direction Proposal Module*. With the optimal retrieval direction, the *Dynamics Predictor* predicts the support relations between each object adjacent to $O_i$.

where $p$ means the distribution of $z_i$ and $G$ means the Gaussian distribution.

To further select the optimal retrieval direction, in the *Direction Scoring Module*, we use Multi-Layer Perceptrons (MLPs) called $F_{DSM}$ to take the feature concatenation of $f_{O_i}$ and $f_S$ and $d_i$, and predict the corresponding movement score $\hat{m}_i$. We employ the L1-loss between $\hat{m}_i$ and the ground truth movement $m_i$ as the loss:

$$L_{score} = |m_i - F_{DSM}(f_{O_i}, f_S, d_i)|. \tag{3}$$

Such combination of Direction Proposal and Direction Scoring can progressively model the distribution of promising direction candidates and select the optimal retrieval direction leveraging the capability of two different-structured networks, which is much better than directly proposing an action direction, as promising actions lie in a distribution of multi-modal and dissimilar candidates but with similar scores, while a single direction prediction network could only generate single-modal predictions.

### D. Local Dynamics Predictor

With the proposed optimal direction $d_k$ for an object $O_i$, the next step is to estimate the resulting dynamics of other objects when $d_k$ is applied on $O_i$, *i.e.*, whether these objects will transition into a new state or remain static. However, due to the intricate relations between each objects within the clutter, directly predicting the dynamics state of each object in the clutter using a single network is both unfeasible and unnecessary. For example, the relations between two distant objects can be transferred by objects in between, which is quite a complicated process.

To address this challenge, we start from training *Local Dynamics Predictor*, aiming to estimate the adjacent object's dynamics state when a particular action is applied, as thus relations are much easier for the model to make accurate inference.

Inspired by the particle-based dynamics predictor [4], which uses particles to represent objects has demonstrated great performance in modelling the dynamics of adjacent objects, we use particle-based predictor to estimate the dynamics states of any object $O_j$ when retrieval is applied at direction $d_k$ on $O_i$, where $O_j$ is adjacent to $O_i$. To be specific, to extract the feature $g_{O_i}$ of each object $O_i$, instead of directing encoding $O_i$ into a global feature, we sample $r$ ($r = 256$ in our paper) particles on $O_i$ and $O_j$ using farthest-point sampling, and merge the two groups of particles together with $d_k$ as the additional channel for $O_i$. Then we use Segmentation-version PointNet++ to extract per-point features of particles hierarchically and thus use the averaged features of sampled particles as the dynamics feature $g_{ij,k}$. Then, we use a Multi-layer Perceptron (MLP) called $F_{LDP}$ to take the dynamics feature $g_{ij,k}$ as input, to predict the dynamics state of $O_j$. To train this module, we employ Binary Cross Entropy (BCE) loss with ground truth labelled as 0 or 1 representing its dynamics state:

$$L_{dynamic} = BCE(Label, F_{LDP}(g_{ij,k})). \tag{4}$$

The **Experiments** section will show the empirical performance increase by using the particle-based dynamics model instead of directly extracting the object representations.

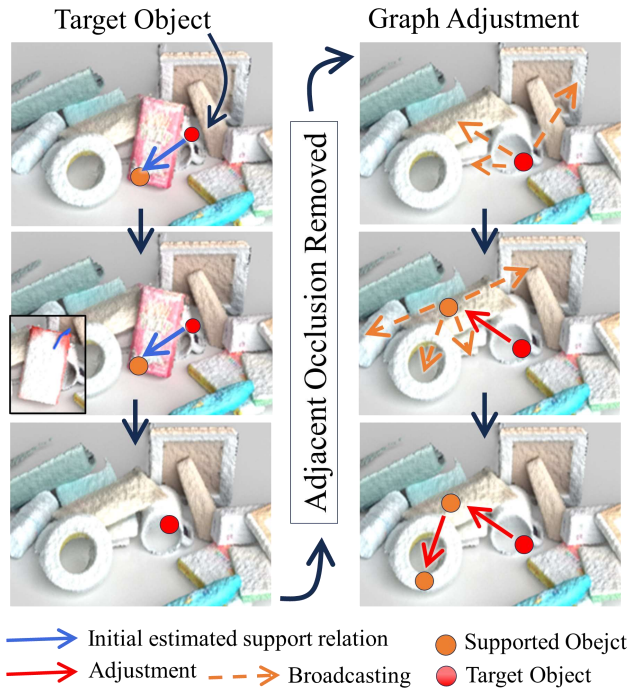


Fig. 3: **Graph Adjustment** when the occlusion pink box is removed and the system re-broadcasts the supporting relations from the mug to be retrieved (target object).

### *E. Clutter Solver: Recursive Support Relation Broadcasting*

With the trained *Retrieval Direction Predictor* and *Local Dynamics Predictor*, for any object $O_i$ in *Support Graph*, we can predict whether retrieving $O_i$ at the optimal retrieval direction will lead to the movement of its adjacent objects. Specifically, we use *Retrieval Direction Predictor* to propose the optimal retrieval direction $d_k$. Subsequently, for each $O_j \in N(O_i)$ (where $N(O_i)$ is the set of objects adjacent to $O_i$) we process them through *Local Dynamics Predictor* respectively and obtain $m_{i,j}$ represent the movement of $O_j$ under the force of $O_i$ at its optimal retrieval direction $d_k$. We call this process the *broadcast* of the dynamics of $O_i$. For $m_{i,j}$ larger than a threshold $th_m$, $O_j$ can be regarded as supported by $O_i$ and also the child node of $O_i$ in *Support Graph*. It means to retrieve $O_i$ without collision, $O_j$ must be retrieved in advance. Besides, under the assumption that there is no mutually supportive relationship (such situation is quite rare and will require two robot arms to ensure the manipulation safety), we can ignore $O_i$'s parent nodes among $N(O_i)$ when broadcasting its dynamics.

Starting from target object $O_t$, and then its child nodes, we recursively use this mechanism to explore the *Support Graph* and obtain the final graph after self-convergence. Note that not all support relations in the clutter is estimated because many of them are irrelevant to our goal of target retrieving. With *clutter Solver* we can largely improve the computing efficiency in maximum extend than querying all support relations in the clutter in turn.

As the whole scene is partially observed, some objects which should be included in *Support Graph* $\mathcal{G}$ may be ignored because the influence of occlusions. Specifically, they are actually supported by the target object, but currently occluded by other objects and thus temporarily could not exist in $\mathcal{G}$. However, when the occlusions are removed, the supporting relation between the target and this object should be revealed. To eliminate the impact of this case, we propose the *Graph Adjustment* process to enhance the robustness for such a long-horizon task with occlusions. Specifically, when an object $O_i$ is the next object to be retrieved, we conduct the broadcast again from $O_i$. When all the objects supported by $O_i$ all exist in $\mathcal{G}$, there is no need to further update $\mathcal{G}$. However, when there are novel objects supported by $O_i$ not existing in $\mathcal{G}$, $\mathcal{G}$ should be recursively updated from the node of $O_i$. Figure 3 gives a demonstration of this process. The system originally estimates that only the pink box is supported by the target mug. when the pink box is removed, before retrieving the mug, the system first estimates the local dynamics of the mug and finds the previously hidden envelop is also support by the mug. Then, the system further broadcasts the support relations from the envelop and recursively finds more support relations.

The benefit of *Graph Adjustment* lies in two aspects. First, it can help to identify objects supported by $O_i$ which are wrongly inferred as non-supported objects before manipulation. Because as the occluding objects is retrieved, more details of the hidden object are exposed and help to give a more accurate dynamics estimation. Compared to inferring $\mathcal{G}$ from scratch (*i.e.*, the target object), broadcasting the graph from $O_i$ only when a new object will exist in the $\mathcal{G}$ is much more cost efficient.

With the constructed *Support Graph* $\mathcal{G}_{t_i}$ at each time step $t_i$, only the objects in $\mathcal{G}_{t_i}$ should be considered to retrieve, while other objects don't have support relations with the target. We can easily find one object $O_{t_i}$ in $\mathcal{G}_{t_i}$ that is directly or indirectly supported by the target object while not supporting other objects (*i.e.*, the node $O_{t_i}$ with no outdegree in $\mathcal{G}_{t_i}$), and retrieve it away by estimating the manipulation affordance.

### *F. Manipulation Affordance Predictor*

After determining a feasible retrieval sequence through the generated *Support Graph* and selecting the object $O_{k_i}$ to be manipulated at the current time step, our objective is to grasp this object while minimizing displacements or collisions with adjacent objects during manipulation. Inspired by previous works [61, 68, 64, 81, 9, 63, 26] that demonstrated efficacy of visual affordance in offering generalizable actionable priors for diverse objects in 3D manipulation scenarios, we introduce the *Manipulation Affordance Predictor* module. This module is designed to propose the optimal grasp point $p_i \in \mathbb{R}^3$ and grasp direction $r_i \in SO(3)$ to safely manipulate the object $O_{k_i}$ without disturbing others.

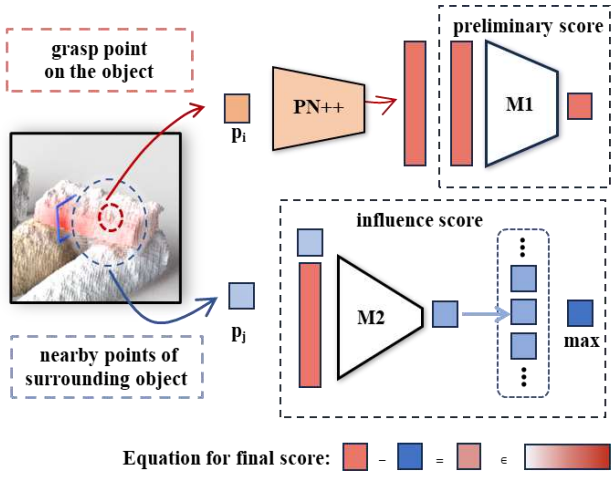


Fig. 4: **Affordance Scoring Module**. To estimate the affordance score for a grasp point, we first calculate the preliminary affordance score solely based on the point itself, and then we evaluate the influence score by estimating the potential impact. The final affordance score is obtained by subtracting this influence score from the preliminary score.

The *Manipulation Affordance Predictor* consists of two submodules: the *Affordance Scoring Module* and the *Grasp Direction Predictor Module*. The *Affordance Scoring Module* predicts the affordance score for difference grasp points, enabling the identification of high-quality grasp points and the selection of an optimal grasp point $p_i$ on the object. Then, the *Grasp Direction Predictor Module* generates the optimal grasp direction $r_i$ for the selected grasp point.

In the *Affordance Scoring Module* (Figure 4), the estimation of affordance scores for various points necessitates a comprehensive consideration of both the object's geometry and its surrounding environment. Initially, we calculate the preliminary affordance score for the grasp point $p_i$, excluding the influence of the surrounding clutter. Subsequently, we evaluate the potential impact of $p_i$ on the surrounding clutter, refining the preliminary affordance score. The final affordance score is then obtained by subtracting this refined impact from the preliminary affordance score.

To start with, we leverage PointNet++ [44] to process the partial point cloud, extracting per-point features denoted as $f_{p_i}$. Then we use an MLP $M_1$ to decode the preliminary affordance score for each point based on its feature. To estimate the potential influence a grasp point might exert on other objects, we identify all nearby points of $p_i$ from different objects, forming a set $A_{p_i}$, where for any point $p_j \in A_{p_i}$, $dist(p_i, p_j) < \epsilon_x$ and $p_j \notin O_{k_i}$. Then, we evaluate the potential impact of a grasp point $p_i$ on other object, by using an MLP $M_2$ to process both the grasp point feature $f_{p_i}$ and the position of the nearby point $p_j$ to decode the influence score. The final affordance score $s_p^i$ of $p_i$ is determined by subtracting the maximum influence score from the preliminary affordance score:

$$s_p^i = M_1(f_{p_i}) - \max_{i \in A_{p_i}} M_2(f_{p_i}, p_j). \tag{5}$$

In the *Grasp Pose Predictor Module*, similar to the method utilized in the *Retrieval Direction Predictor* (Section IV-C), we employ a conditional Variational Autoencoder (cVAE) to generate multiple candidates for grasp pose, and use Multi-Layer Perceptrons (MLPs) to predict the corresponding movement caused by these pose candidates. The final grasp pose is determined by selecting the poses which enable the robot manipulator to grasp and retrieve the object successfully without collision. The intricate architectural design and training strategy closely resemble those employed in the *Retrieval Direction Predictor*, as outlined in Section IV-C.

## V. Experiments

### A. Setup

For **simulation environment**, we equip OMNIVERSE ISAAC SIM [25] with 1 Franka Panda robot arm and 16 categories of thousands of different objects from ShapeNet [3], building up 4 different and realistic scenarios: kitchen, desk, food and sundries. The detailed data statistics is shown in section 1.1 in the appendix. To train the manipulation policy, we generate 5,000 different complicated clutters with different combinations of diverse objects for each scenario, where each clutter contains 15 objects on average. This simulation environment with proposed dataset contains much more diverse objects and complicated clutters compared with previous evaluation environments [33, 16, 66].

For **real-world setup**, we build the 4 scenarios (representative cases shown in the Figure 5), use Microsoft Azure Kinect (which has demonstrated high-precision with slight noises for robotic manipulation [38]) to capture the point cloud of target scenes, and Robot Operating System (ROS) [46] to control the Franka Panda robot arm for manipulation. To capture the point cloud of each object, we use Segment Anything (SAM) [18] to segment each object, and project the corresponding depth image to the point cloud.

### B. Baseline, Ablations and Metrics

To demonstrate the superiority of our framework in cluttered scenarios, we compare with 2 most recent strong **baselines**:

- **RD-GNN** [16] that directly classifies objects relations in the scene using Graph Neural Networks (GNN).
- **SafePicking** [56] that uses object-level mapping and learning-based motion planning to achieve safe object retrieval.

To demonstrate the effectiveness of different components of our framework, we compare with 4 **ablated versions**:

- **Ours w/o DP** that removes the Retrieval Direction Predictor (DP) and always takes the upward direction for manipulation;
- **Ours w/o PR** that replaces the Particle-Based Representation (PR) with Object-level Representation for local dynamics prediction;

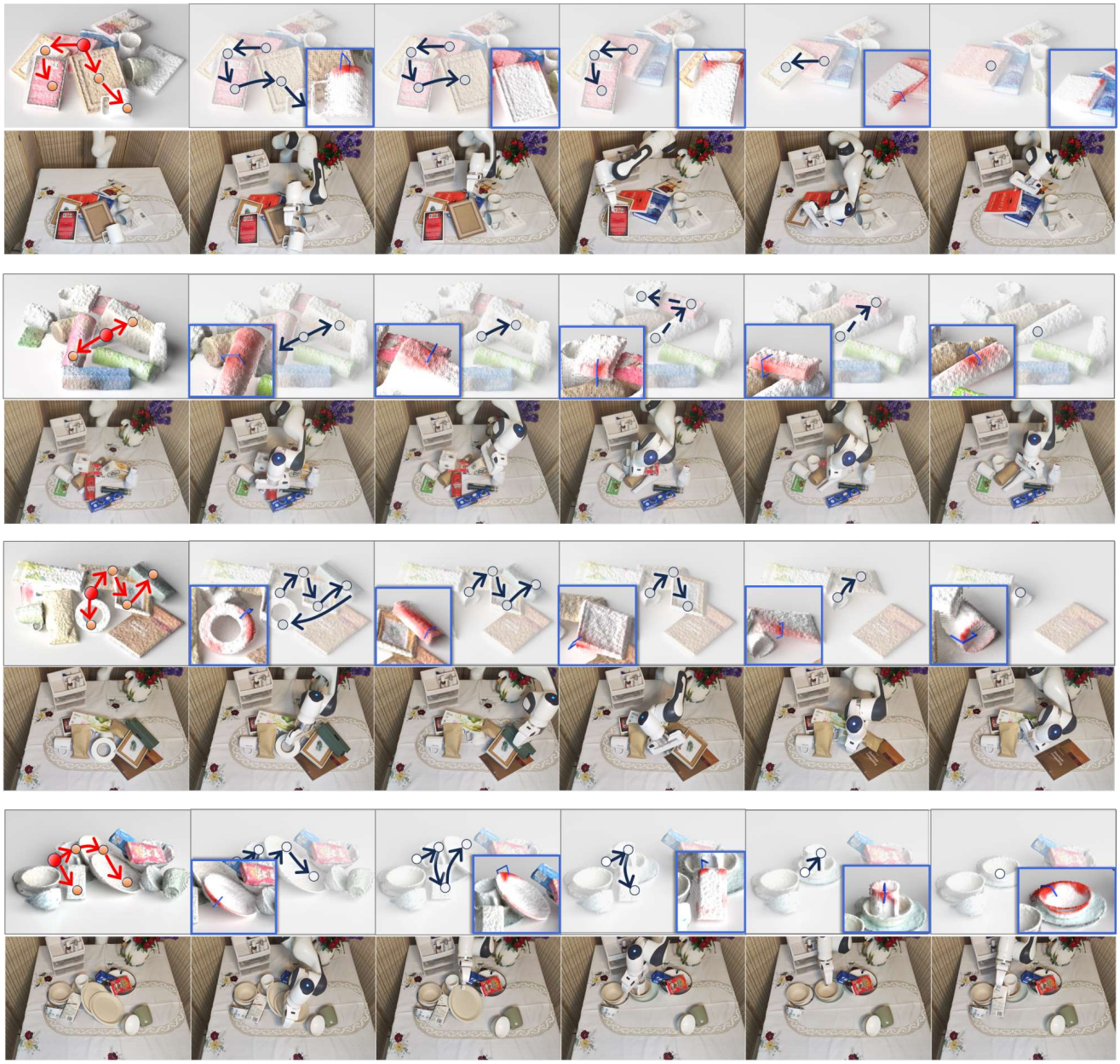

Fig. 5: **Manipulation Sequence for Real-World Clutters with Captured Point Clouds.** We show the 4 cases respectively demonstrating the desk, food, sundries and kitchen scenarios. The second case contains occlusion removal and thus executes the **Graph Adjustment** process after moving away the white box in column 3.

- **Ours w/o RB** that removes the Recursive Broadcasting (RB) process, and directly predicts the support relation between each 2 object;
- **Ours w/o GA** that removes the Graph Adjustment (GA) process and just uses the initial support map to conduct retrieval, without checking support relation after each manipulation step.

For evaluation, we employ the following **metrics**:

- **Retrieval Success Rate** that evaluates whether any displacement occurs during manipulation.
- **Accumulated Displacement Distance** demonstrates the mean of accumulated displacement distance for each scene in the test set during manipulation.
- **Retrieval Steps** that counts the average steps of retrieval under same successful retrieval cases.
- **Relation Prediction Success Rate** reflects whether the proposed dynamics states are correct.
- **Retrieval Direction Success Rate** that evaluates whether the proposed retrieval direction will lead to the least displacements of nearby objects.

### *C. Results and Analysis*

Table I, II and III show the large-scale evaluation results on the 3 main evaluation metrics in simulation. For each scenario, we set 357 different clutters for evaluation. Our framework

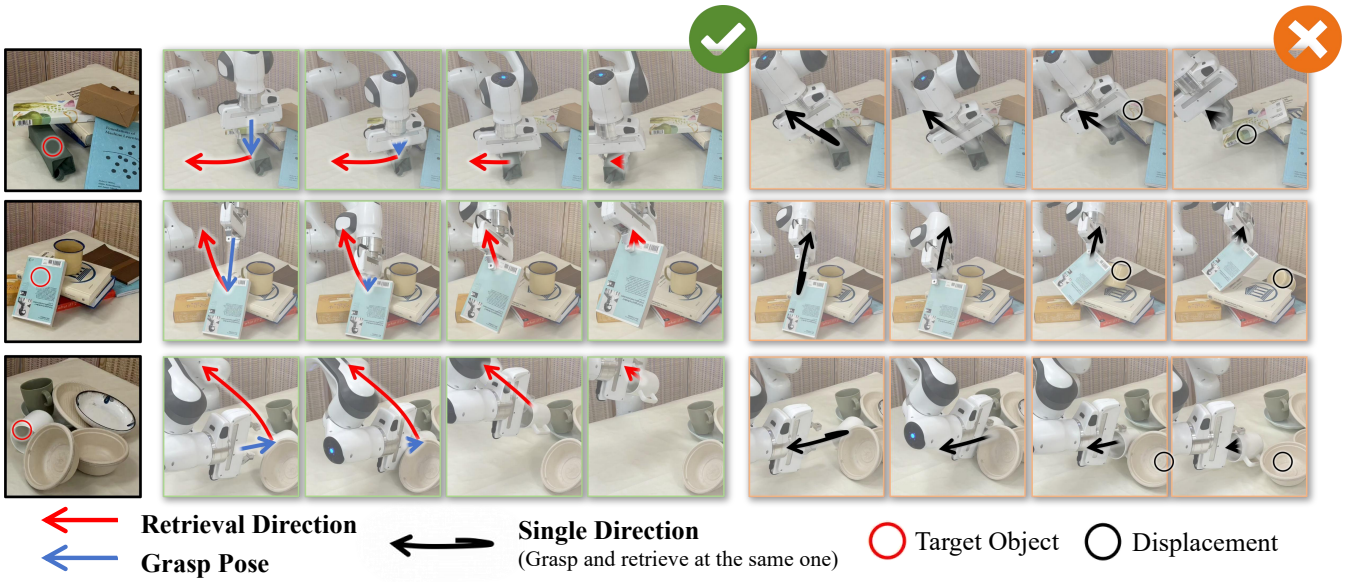


Fig. 6: **Clarification between Grasp Pose and Retrieval Direction.** This figure reflects the manipulation flexibility brought by the separated **grasp pose** and **retrieval direction**. For example, the last row shows that the differentiation between **grasp pose** and **retrieval direction** enables the manipulator to reach optimality both in **grasp stability** and **collision avoidance with other objects** while the **single direction** easily leads to collision.

TABLE I: **Clutter Object Retrieval Success Rate** which show overall success rate of our algorithm. It demonstrates that our design not only outperforms other algorithms but each part is crucial.

| Method | Kitchen | Desk | Food | Sundries |
|---|---|---|---|---|
| Ours w/o DP | 0.72±0.038 | 0.79±0.041 | 0.73±0.032 | 0.81±0.037 |
| Ours w/o GA | 0.64±0.046 | 0.66±0.041 | 0.62±0.037 | 0.69±0.045 |
| Ours w/o RB | 0.23±0.080 | 0.30±0.074 | 0.19±0.079 | 0.33±0.081 |
| Ours w/o PR | 0.65±0.043 | 0.68±0.046 | 0.64±0.045 | 0.66±0.039 |
| RD-GNN | 0.30±0.086 | 0.33±0.085 | 0.25±0.097 | 0.25±0.091 |
| SafePicking | 0.37±0.085 | 0.42±0.078 | 0.34±0.081 | 0.35±0.077 |
| **Ours** | **0.79±0.024** | **0.84±0.028** | **0.76±0.029** | **0.83±0.026** |

TABLE II: **Retrieval Steps** under same successful cases. This criterion show that our algorithm achieves higher grasping efficiency compared to other baseline algorithms. Fewer number of grasping steps indicate higher efficiency.

| method | Kitchen | Desk | Food | Sundries |
|---|---|---|---|---|
| Ours w/o DP | 5.21±0.66 | 4.75±0.71 | 5.13±0.62 | 5.62±0.67 |
| Ours w/o GA | 4.31±0.84 | 5.02±0.69 | 5.29±0.75 | 4.96±0.79 |
| Ours w/o RB | 7.95±1.13 | 7.72±1.09 | 7.91±1.21 | 7.75±1.08 |
| Ours w/o PR | 5.08±1.08 | 4.86±1.11 | 5.72±1.07 | 5.4±1.14 |
| RD-GNN | 7.41±1.38 | 7.83±1.27 | 7.56±1.32 | 6.67±1.43 |
| SafePicking | 7.85±1.52 | 7.57±1.49 | 7.73±1.45 | 7.24±1.56 |
| **Ours** | **4.11±0.52** | **4.61±0.43** | **4.95±0.54** | **4.51±0.49** |

can infer the support graph of the scene, move away objects directly or indirectly supported by the target object step by step, and finally retrieve the target object safely.

TABLE III: **Accumulated Displacement Distance** shows the accumulated displacement distance for each scene (unit: centimeter). Lower total displacement Distance indicates the safer manipulation process.

| method | Kitchen | Desk | Food | Sundries |
|---|---|---|---|---|
| Ours w/o DP | 9.44±0.86 | 8.81±0.94 | 9.63±0.0.90 | 9.58±0.91 |
| Ours w/o GA | 19.21±1.14 | 16.67±1.27 | 17.15±1.08 | 18.70±1.19 |
| Ours w/o RB | 27.33±2.74 | 24.96±2.45 | 28.21±3.07 | 25.78±2.51 |
| Ours w/o PR | 18.24±1.02 | 14.09±0.95 | 16.93±1.24 | 17.71±1.16 |
| RD-GNN | 23.49±1.71 | 23.57±1.96 | 26.78±1.85 | 28.14±1.43 |
| SafePicking | 21.54±1.68 | 17.69±1.72 | 23.26±1.42 | 22.76±1.59 |
| **Ours** | **6.57±0.77** | **5.89±0.64** | **6.21±0.79** | **6.04±0.73** |

The rapid performance increase from **RD-GNN**, **SafePicking**, **Ours w/o RB** (which all directly predict relations between each 2 objects) to **Our Whole Framework** in all the tables demonstrate that, our main design, **Recursive Broadcast**, is the fundamental mechanism for retrieving object in clutter, as the support relation construction capability of the whole scene can be largely boosted by gradually broadcasting the more accurate local dynamics predictions. It is worth mentioning, the second case shows the **Graph Adjustment** case. Specifically, while initially the framework cannot inference the hidden pink box is supported by the target object, after retrieving the white box in the right, the pink box is not occluded and our framework efficiently adjusts that it is supported by the target object, and should be moved away. In this case, the ablated version **Ours w/o GA** will not work, revealed in its performance decrease Table I and II.

For **Retrieval Direction**, Figure 6 and Figure 7 shows that our framework can effectively discriminate manipulation directions based on their potential collisions with other objects. Besides, the comparison between **Ours** and **Ours w/o DP** in Table IV demonstrates that our method with the *Retrieval Direction Predictor* will generate promising manipulation directions and thus leads to better performance.

TABLE IV: **Retrieval Direction Success Rate** demonstrates the significance of the *Retrieval Direction Predictor*.

| method | kitchen | desk | food | sundries |
|---|---|---|---|---|
| Ours w/o DP (no std) | 0.62 | 0.68 | 0.63 | 0.65 |
| **Ours** | **0.89±0.021** | **0.92±0.019** | **0.87±0.022** | **0.90±0.027** |

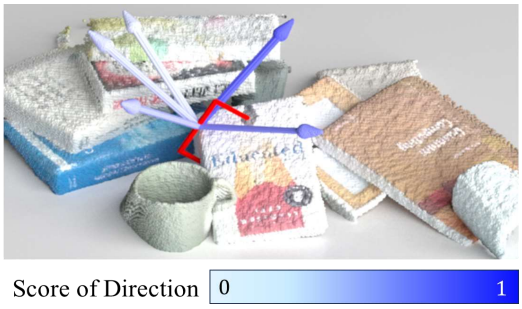


Fig. 7: **Scores of different directions** demonstrate that our framework proposes action directions that cause minimal disturbance to other objects. Higher scores mean better directions.

For **Dynamics Prediction**, as shown in the comparison between **Ours** and **Ours w/o PR** in Table V, the particle-based representation, which is employed in our framework, can highly improve the prediction accuracy compared with the object-level representation [4].

TABLE V: **Relation Prediction Success Rate** shows the accurate dynamics prediction of *Local Dynamics Predictor*.

| method | Kitchen | Desk | Food | Sundries |
|---|---|---|---|---|
| Ours w/o PR | 0.73±0.052 | 0.78±0.050 | 0.74±0.053 | 0.77±0.058 |
| **Ours** | **0.89±0.029** | **0.93±0.028** | **0.90±0.031** | **0.91±0.028** |

Besides evaluating in simulator, we also conduct real world experiments. Table VI shows the success rate in the real world scenarios. In each scenario, we set 10 or 15 different clutters and execute the policy.

TABLE VI: **Retrieval Success Rate** in the real world demonstrates the superior performance of our method compared to the baselines in real world clutter object retrieval tasks.

| method | kitchen | desk | food | sundries |
|---|---|---|---|---|
| RD-GNN | 2/10 | 4/15 | 6/15 | 3/10 |
| SafePicking | 4/10 | 7/15 | 8/15 | 4/10 |
| **Ours** | **8/10** | **11/15** | **10/15** | **7/10** |

## VI. Conclusion

In this paper, we study the problem of cluttered objects manipulation. As the clutter could be complicated and it is highly difficult to directly infer the support relations between any 2 objects, we propose a framework broadcasting support relations recursively from local dynamics, to effectively and efficiently predict the support graph of the whole clutter, guiding the safe retrieval of the target object. Extensive experiments showcase the superiority of our proposed framework.

## VII. Acknowledge

This paper is supported by National Natural Science Foundation of China - General Program (62376006), National Natural Science Foundation of China (No. 62136001), The National Youth Talent Support Program (8200800081).

## References

[1] Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021.

[2] Michel Breyer, Jen Jen Chung, Lionel Ott, Roland Siegwart, and Juan Nieto. Volumetric grasping network: Real-time 6 dof grasp detection in clutter. In *CoRL*, pages 1602–1611. PMLR, 2021.

[3] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[4] Haonan Chen, Yilong Niu, Kaiwen Hong, Shuijing Liu, Yixuan Wang, Yunzhu Li, and Katherine Rose Driggs-Campbell. Predicting object interactions with behavior primitives: An application in stowing tasks. In *CoRL*, pages 358–373. PMLR, 2023.

[5] Siang Chen, Wei Tang, Pengwei Xie, Wenming Yang, and Guijin Wang. Efficient heatmap-guided 6-dof grasp detection in cluttered scenes. *IEEE Robotics and Automation Letters*, 2023.

[6] Sang Hun Cheong, Brian Y Cho, Jinhwi Lee, ChangHwan Kim, and Changjoo Nam. Where to relocate?: Object rearrangement inside cluttered and confined environments for robotic manipulation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7791–7797. IEEE, 2020.

[7] Qiyu Dai, Yan Zhu, Yiran Geng, Ciyu Ruan, Jiazhao Zhang, and He Wang. Graspnerf: Multiview-based 6-dof grasp detection for transparent and specular objects using generalizable nerf. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1757–1763. IEEE, 2023.

[8] Michael Danielczuk, Andrey Kurenkov, Ashwin Balakrishna, Matthew Matl, David Wang, Roberto Martín-Martín, Animesh Garg, Silvio Savarese, and Ken Gold-

berg. Mechanical search: Multi-step retrieval of a target object occluded by clutter. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 1614–1621. IEEE, 2019.

[9] Kairui Ding, Boyuan Chen, Ruihai Wu, Yuyang Li, Zongzheng Zhang, Huan-ang Gao, Siqi Li, Yixin Zhu, Guyue Zhou, Hao Dong, et al. Preafford: Universal affordance-based pre-grasping for diverse objects and environments. *arXiv preprint arXiv:2404.03634*, 2024.

[10] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11444–11453, 2020.

[11] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *ICRA*, pages 2786–2793. IEEE, 2017.

[12] Ankit Goyal, Arsalan Mousavian, Chris Paxton, Yu-Wei Chao, Brian Okorn, Jia Deng, and Dieter Fox. Ifor: Iterative flow minimization for robotic object rearrangement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14787–14797, 2022.

[13] Ruiqi Guo and Derek Hoiem. Support surface prediction in indoor scenes. In *ICCV*, pages 2144–2151, 2013.

[14] Baichuan Huang, Shuai D Han, Jingjin Yu, and Abdeslam Boularias. Visual foresight trees for object retrieval from clutter with nonprehensile rearrangement. *IEEE Robotics and Automation Letters*, 7(1):231–238, 2021.

[15] Shi-Sheng Huang, Hongbo Fu, Ling-Yu Wei, and Shi-Min Hu. Support substructures: Support-induced part-level structural representation. *IEEE transactions on visualization and computer graphics*, 22(8):2024–2036, 2015.

[16] Yixuan Huang, Adam Conkey, and Tucker Hermans. Planning for multi-object manipulation with graph neural network relational classifiers. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1822–1829. IEEE, 2023.

[17] Rainer Kartmann, Fabian Paus, Markus Grotz, and Tamim Asfour. Extraction of physically plausible support relations to predict and validate manipulation action effects. *IEEE Robotics and Automation Letters*, 3(4): 3991–3998, 2018.

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[19] Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B Tenenbaum. Learning to act from actionless videos through dense correspondences. *arXiv preprint arXiv:2310.08576*, 2023.

[20] Andrey Kurenkov, Joseph Taglic, Rohun Kulkarni, Marcus Dominguez-Kuhne, Animesh Garg, Roberto Martín-Martín, and Silvio Savarese. Visuomotor mechanical search: Learning to retrieve target objects in clutter. In *IROS*, pages 8408–8414. IEEE, 2020.

[21] Jinhwi Lee, Younggil Cho, Changjoo Nam, Jonghyeon Park, and Changhwan Kim. Efficient obstacle rearrangement for object manipulation tasks in cluttered environments. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 183–189. IEEE, 2019.

[22] Dayou Li, Pengkun Wei, Chenkun Zhao, Shuo Yang, Yibin Li, and Wei Zhang. A mobile manipulation system for automated replenishment in the field of unmanned retail. In *2023 IEEE International Conference on Mechatronics and Automation (ICMA)*, pages 644–649. IEEE, 2023.

[23] Yu Li, Xiaojie Zhang, Ruihai Wu, Zilong Zhang, Yiran Geng, Hao Dong, and Zhaofeng He. Unidoormanip: Learning universal door manipulation policy over large-scale and diverse door manipulation environments. *arXiv preprint arXiv:2403.02604*, 2024.

[24] Yunzhu Li, Jiajun Wu, Russ Tedrake, Joshua B Tenenbaum, and Antonio Torralba. Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. *arXiv preprint arXiv:1810.01566*, 2018.

[25] Jacky Liang, Viktor Makoviychuk, Ankur Handa, Nuttapong Chentanez, Miles Macklin, and Dieter Fox. Gpu-accelerated robotic simulation for distributed reinforcement learning. In *CoRL*, pages 270–282. PMLR, 2018.

[26] Suhan Ling, Yian Wang, Shiguang Wu, Yuzheng Zhuang, Tianyi Xu, Yu Li, Chang Liu, and Hao Dong. Articulated object manipulation with coarse-to-fine affordance for mitigating the effect of point cloud noise. *ICRA*, 2024.

[27] Chenchen Liu, Yang Jin, Kehan Xu, Guoqiang Gong, and Yadong Mu. Beyond short-term snippet: Video relation detection with spatio-temporal global context. In *CVPR*, pages 10840–10849, 2020.

[28] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, pages 852–869. Springer, 2016.

[29] Haoran Lu, Yitong Li, Ruihai Wu, Chuanruo Ning, Yan Shen, and Hao Dong. Unigarment: A unified simulation and benchmark for garment manipulation. *ICRA Workshop on Deformable Object Manipulation*, 2024.

[30] Yuhao Lu, Yixuan Fan, Beixing Deng, Fangfu Liu, Yali Li, and Shengjin Wang. Vl-grasp: a 6-dof interactive grasp policy for language-oriented objects in cluttered indoor scenes. In *IROS*, pages 976–983. IEEE, 2023.

[31] Xiao Ma, David Hsu, and Wee Sun Lee. Learning latent graph dynamics for deformable object manipulation. *arXiv preprint arXiv:2104.12149*, 2, 2021.

[32] Rasoul Mojtahedzadeh, Abdelbaki Bouguerra, Erik Schaffernicht, and Achim J Lilienthal. Support relation analysis and decision making for safe robotic manipulation tasks. *Robotics and Autonomous Systems*, 71:99–117, 2015.

[33] Tomohiro Motoda, Damien Petit, Takao Nishi, Kazuyuki Nagata, Weiwei Wan, and Kensuke Harada. Multi-step object extraction planning from clutter based on support

relations. *IEEE Access*, 2023.

[34] Peiyuan Ni, Wenguang Zhang, Xiaoxiao Zhu, and Qixin Cao. Pointnet++ grasping: Learning an end-to-end spatial grasp generation algorithm from sparse point clouds. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3619–3625. IEEE, 2020.

[35] Yinyu Nie, Jian Chang, Ehtzaz Chaudhry, Shihui Guo, Andi Smart, and Jian Jun Zhang. Semantic modeling of indoor scenes with support inference from a single photograph. *Computer Animation and Virtual Worlds*, 29(3-4):e1825, 2018.

[36] Yinyu Nie, Shihui Guo, Jian Chang, Xiaoguang Han, Jiahui Huang, Shi-Min Hu, and Jian Jun Zhang. Shallow2deep: Indoor scene modeling by single image understanding. *Pattern Recognition*, 103:107271, 2020.

[37] Yinyu Nie, Jian Chang, and Jian Jun Zhang. Content-aware semantic indoor scene modeling from a single image. In *Intelligent Scene Modeling and Human-Computer Interaction*, pages 129–145. Springer, 2021.

[38] Chuanruo Ning, Ruihai Wu, Haoran Lu, Kaichun Mo, and Hao Dong. Where2explore: Few-shot affordance learning for unseen novel categories of articulated objects. *arXiv preprint arXiv:2309.07473*, 2023.

[39] Swagatika Panda, AH Abdul Hafez, and CV Jawahar. Learning semantic interaction among graspable objects. In *Pattern Recognition and Machine Intelligence: 5th International Conference, PReMI 2013, Kolkata, India, December 10-14, 2013. Proceedings 5*, pages 304–312. Springer, 2013.

[40] Swagatika Panda, AH Abdul Hafez, and CV Jawahar. Learning support order for manipulation in clutter. In *2013 IEEE/RSJ international conference on intelligent robots and systems*, pages 809–815. IEEE, 2013.

[41] Swagatika Panda, Abdul Hafez Abdul Hafez, and CV Jawahar. Single and multiple view support order prediction in clutter for manipulation. *Journal of Intelligent & Robotic Systems*, 83:179–203, 2016.

[42] Fabian Paus and Tamim Asfour. Probabilistic representation of objects and their support relations. In *Experimental Robotics: The 17th International Symposium*, pages 510–519. Springer, 2021.

[43] Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W Battaglia. Learning mesh-based simulation with graph networks. *arXiv preprint arXiv:2010.03409*, 2020.

[44] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[45] Yuzhe Qin, Rui Chen, Hao Zhu, Meng Song, Jing Xu, and Hao Su. S4g: Amodal single-view single-shot se (3) grasp detection in cluttered scenes. In *CoRL*, pages 53–65. PMLR, 2020.

[46] Morgan Quigley, Josh Faust, Tully Foote, Jeremy Leibs, et al. Ros: an open-source robot operating system.

[47] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International conference on machine learning*, pages 8459–8468. PMLR, 2020.

[48] Max Schwarz, Anton Milan, Arul Selvam Periyasamy, and Sven Behnke. Rgb-d object detection and semantic segmentation for autonomous manipulation in clutter. *The International Journal of Robotics Research*, 37(4-5): 437–451, 2018.

[49] Haochen Shi, Huazhe Xu, Zhiao Huang, Yunzhu Li, and Jiajun Wu. Robocraft: Learning to see, simulate, and shape elasto-plastic objects with graph networks. *arXiv preprint arXiv:2205.02909*, 2022.

[50] Haochen Shi, Huazhe Xu, Samuel Clarke, Yunzhu Li, and Jiajun Wu. Robocook: Long-horizon elasto-plastic object manipulation with diverse tools. *arXiv preprint arXiv:2306.14447*, 2023.

[51] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012.

[52] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

[53] HJ Terry Suh and Russ Tedrake. The surprising effectiveness of linear models for visual foresight in object pile manipulation. In *Algorithmic Foundations of Robotics XIV: Proceedings of the Fourteenth Workshop on the Algorithmic Foundations of Robotics 14*, pages 347–363. Springer, 2021.

[54] Martin Sundermeyer, Arsalan Mousavian, Rudolph Triebel, and Dieter Fox. Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13438–13444. IEEE, 2021.

[55] Bingjie Tang and Gaurav S Sukhatme. Selective object rearrangement in clutter. In *CoRL*, pages 1001–1010. PMLR, 2023.

[56] Kentaro Wada, Stephen James, and Andrew J. Davison. Safepicking: Learning safe object extraction via object-level mapping. *2022 International Conference on Robotics and Automation (ICRA)*, pages 10202–10208, 2022. URL https://api.semanticscholar.org/CorpusID:246823630.

[57] Chenxi Wang, Hao-Shu Fang, Minghao Gou, Hongjie Fang, Jin Gao, and Cewu Lu. Graspness discovery in clutters for fast and accurate grasp detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15964–15973, 2021.

[58] Yian Wang, Ruihai Wu, Kaichun Mo, Jiaqi Ke, Qingnan Fan, Leonidas Guibas, and Hao Dong. Adaafford: Learning to adapt manipulation affordance for 3d articulated objects via few-shot interactions. *European conference*

*on computer vision (ECCV 2022)*, 2022.

[59] Yixuan Wang, Yunzhu Li, Katherine Driggs-Campbell, Li Fei-Fei, and Jiajun Wu. Dynamic-resolution model learning for object pile manipulation. *RSS*, 2023.

[60] Ruihai Wu, Kehan Xu, Chenchen Liu, Nan Zhuang, and Yadong Mu. Localize, assemble, and predicate: Contextual object proposal embedding for visual relation detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12297–12304, 2020.

[61] Ruihai Wu, Yan Zhao, Kaichun Mo, Zizheng Guo, Yian Wang, Tianhao Wu, Qingnan Fan, Xuelin Chen, Leonidas Guibas, and Hao Dong. Vat-mart: Learning visual action trajectory proposals for manipulating 3d articulated objects. *ICLR*, 2021.

[62] Ruihai Wu, Kai Cheng, Yan Zhao, Chuanruo Ning, Guanqi Zhan, and Hao Dong. Learning environment-aware affordance for 3d articulated object manipulation under occlusions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Re2NHYoZ5l.

[63] Ruihai Wu, Chuanruo Ning, and Hao Dong. Learning foresightful dense visual affordance for deformable object manipulation. In *IEEE International Conference on Computer Vision (ICCV)*, 2023.

[64] Ruihai Wu, Haoran Lu, Yiyan Wang, Yubo Wang, and Dong Hao. Unigarmentmanip: A unified framework for category-level garment manipulation via dense visual correspondence. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[65] Annie Xie, Frederik Ebert, Sergey Levine, and Chelsea Finn. Improvisation through physical understanding: Using novel objects as tools with visual foresight. *arXiv preprint arXiv:1904.05538*, 2019.

[66] Tao Xiong, Chengchao Yu, Tianyu Xiong, and Huiliang Shang. A geometric inference framework based on physical blocks for ordered grasping robots in clutter. In *2021 International Conference on Networking Systems of AI (INSAI)*, pages 62–68. IEEE, 2021.

[67] Kechun Xu, Shuqi Zhao, Zhongxiang Zhou, Zizhang Li, Huaijin Pi, Yifeng Zhu, Yue Wang, and Rong Xiong. A joint modeling of vision-language-action for target-oriented grasping in clutter. *arXiv preprint arXiv:2302.12610*, 2023.

[68] Ran Xu, Yan Shen, Xiaoqi Li, Ruihai Wu, and Hao Dong. Naturalvlm: Leveraging fine-grained natural language for affordance-guided visual manipulation. *arXiv preprint arXiv:2403.08355*, 2024.

[69] Feng Xue, Shan Xu, Chuan He, Meng Wang, and Richang Hong. Towards efficient support relation extraction from rgbd images. *Information Sciences*, 320: 320–332, 2015.

[70] Shangjie Xue, Shuo Cheng, Pujith Kachana, and Danfei Xu. Neural field dynamics model for granular object piles manipulation. In *CoRL*, pages 2821–2837. PMLR, 2023.

[71] Michael Ying Yang, Wentong Liao, Hanno Ackermann, and Bodo Rosenhahn. On support relations and semantic scene graphs. *ISPRS journal of photogrammetry and remote sensing*, 131:15–25, 2017.

[72] Yang Yang, Hengyue Liang, and Changhyun Choi. A deep learning approach to grasping the invisible. *IEEE Robotics and Automation Letters*, 5(2):2232–2239, 2020.

[73] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

[74] Andy Zeng, Shuran Song, Stefan Welker, Johnny Lee, Alberto Rodriguez, and Thomas Funkhouser. Learning synergies between pushing and grasping with self-supervised deep reinforcement learning. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4238–4245. IEEE, 2018.

[75] Andy Zeng, Shuran Song, Kuan-Ting Yu, Elliott Donlon, Francois R Hogan, Maria Bauza, Daolin Ma, Orion Taylor, Melody Liu, Eudald Romo, et al. 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, 2022.

[76] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. What does stable diffusion know about the 3d scene? *arXiv preprint arXiv:2310.06836*, 2023.

[77] Hanbo Zhang, Yunfan Lu, Cunjun Yu, David Hsu, Xuguang Lan, and Nanning Zheng. INVIGORATE: Interactive Visual Grounding and Grasping in Clutter. In *Proceedings of Robotics: Science and Systems*, Virtual, July 2021. doi: 10.15607/RSS.2021.XVII.020.

[78] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *CVPR*, pages 5532–5540, 2017.

[79] Jiazhao Zhang, Nandiraju Gireesh, Jilong Wang, Xiaomeng Fang, Chaoyi Xu, Weiguang Chen, Liu Dai, and He Wang. Gamma: Graspability-aware mobile manipulation policy learning based on online grasping pose fusion. *arXiv preprint arXiv:2309.15459*, 2023.

[80] Peng Zhang, Xiaoyu Ge, and Jochen Renz. Support relation analysis for objects in multiple view rgb-d images. In *Artificial Intelligence. IJCAI 2019 International Workshops: Macao, China, August 10–12, 2019, Revised Selected Best Papers 28*, pages 41–61. Springer, 2020.

[81] Yan Zhao, Ruihai Wu, Zhehuan Chen, Yourong Zhang, Qingnan Fan, Kaichun Mo, and Hao Dong. Dualafford: Learning collaborative visual affordance for dual-gripper object manipulation. *ICLR*, 2023.

[82] Wei Zhuo, Mathieu Salzmann, Xuming He, and Miaomiao Liu. Indoor scene parsing with instance segmentation, semantic labeling and support relationship inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5429–5437, 2017.