IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Learning from Planned Data to Improve Robotic Pick-and-Place Planning Efficiency

Liang Qin[1], Weiwei Wan[1*], Jun Takahashi[2], Ryo Negishi[2], Masaki Matsushita[2], and Kensuke Harada[1]

***Abstract*—This work proposes a learning method to accelerate robotic pick-and-place planning by predicting shared grasps. Shared grasps are defined as grasp poses feasible to both the initial and goal object configurations in a pick-and-place task. Traditional analytical methods for solving shared grasps evaluate grasp candidates separately, leading to substantial computational overhead as the candidate set grows. To overcome the limitation, we introduce an Energy-Based Model (EBM) that predicts shared grasps by combining the energies of feasible grasps at both object poses. The formulation enables early identification of promising candidates and significantly reduces the search space. Experiments show that our method improves grasp selection performance, offers higher data efficiency, and generalizes well to varying grasps and table heights, given that variations fall within the learned distributions.**

***Index Terms*—Pick-and-place planning, Manipulation.**

I. Introduction

A Pick-and-place task requires choosing grasp poses that satisfy constraints at both the pick and place poses. Typical examples include quality inspection tasks where objects must be placed in specific orientations for evaluation, assembly operations requiring precise positioning and orientation for insertion or alignment, retail scenarios where items must be placed on shelves with logos or labels facing outward, etc. Traditional solutions to grasp selection often relied on reasoning frameworks that evaluated grasp candidates based on Inverse Kinematics (IK), grasp stability, and motion feasibility. However, the computational burden increases rapidly with the number of grasp candidates, as each must be evaluated individually across multiple criteria.

To address this issue, we propose a learning-based approach that predicts grasp candidates considering downstream placement feasibility, thereby enabling efficient and reliable planning under real-world constraints. Our work builds upon the concept of the shared grasp for an object, as introduced in prior studies King et al. [1], Wan and Harada [2], Xu et al. [3]. A shared grasp refers to a grasp pose defined in the object's local coordinate frame that remains feasible (i.e., satisfies collision and IK constraints) under both the initial and goal object poses (Fig. 1). Predicting shared grasps helps narrow down grasp selection, reducing computational overhead in downstream planning. However, direct prediction in the full state space suffers from high dimensionality, requiring a large amount of training data and incurring significant sampling costs. Instead of direct prediction, we propose decomposing the task into two feasible grasp predictions with later merging.

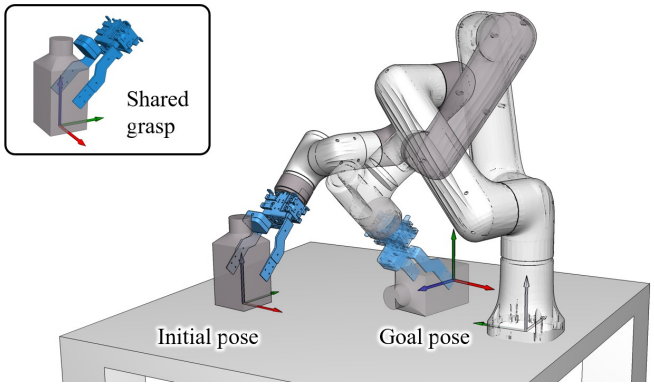

Fig. 1: The blue grasp in the upper-left corner illustrates a candidate feasible at both the initial and goal poses. We refer to such a grasp candidate as a "shared grasp". Our method predicts such grasps to accelerate pick-and-place planning.

We introduce an Energy-Based Model (EBM) that learns the energy landscape over object poses and grasp poses, assigning low energy to those satisfying IK and collision constraints, and high energy to negative ones. Shared grasps are then identified by composing the energies of two feasible grasps. This structured decomposition not only improves planning efficiency but also reduces training complexity. In addition, our method achieves prediction robustness by leveraging pre-annotated grasp candidates in the object's local coordinate system, which ensures the predicted grasp poses dependent on the object pose, avoiding compounded errors that typically arise in methods requiring simultaneous estimation of both object and grasp poses. The pre-annotated grasp candidates make the proposed method well-suited for integration into real-world industrial workflows.

We evaluated the proposed method in experiments and found that it significantly improved grasp selection efficiency while maintaining reliability and success rates. The method also demonstrates generalization to varying grasp and table heights. A supplementary video showcasing the method's integration with visual detection, grasp prediction, and motion planning can be found in the supplementary file.

In summary, our key contributions are as follows:

1) We propose an EBM-based method for shared grasp prediction under tabletop constraints. The prediction can be both used within a controlled and physically feasible space for safety and generalized to unseen grasp poses for adaptivity.
2) We propose a compositional formulation that decomposes

[1]Graduate School of Engineering Science, The University of Osaka, Japan.
[2]H.U. Group Research Institute G.K., Japan.
Contact: Weiwei Wan, wan.weiwei.es@osaka-u.ac.jp

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

the shared grasp prediction into two feasible grasp evaluations. The formulation significantly reduces the complexity of learning in high-dimensional state spaces and better generalizes capacity compared to other methods.

II. Related Work

A. Effective Pick-and-Place Planning

Pick-and-place is a fundamental task in robotic manipulation. It requires precise coordination between grasping and placement [4][5][6]. Such coordination becomes especially important in tasks like assembly [7] and shelf organization [8], where tight spatial interactions are prevalent. However, pick-and-place typically involves a sequence of tasks and motion planning steps that must simultaneously address both symbolic and geometric decisions, making the computation highly complex. One effective strategy is to reduce the search space based on prior knowledge or learned models. For example, Kim et al. [9] introduced a score-space representation that encodes the performance of constraint subspaces and proposed a policy to select promising regions during planning. Yang et al. [10] developed PIGINet, a transformer-based model that predicts the feasibility of symbolic task plans before invoking motion planners. Khodeir et al. [11] proposed a GNN-based relevance model for best-first stream expansion in PDDLStream, which improved search efficiency in large-scale problems.

A key limitation of symbolic task plans is their lack of geometric awareness, and they cannot guarantee that corresponding low-level motions are executable. To address this issue, several methods introduce geometric feasibility predictors based on spatial or visual input. Wells et al. [12] trained an SVM classifier using object position features for box-shaped items. Driess et al. [13] and Xu et al. [14] designed networks that take top-down images to estimate whether a trajectory exists for a given pick or place action. Ait Bouhsain et al. [15], [16] presented AFP-Net and AGFP-Net, which classify the geometric feasibility of discrete pick-and-place actions based on multi-view images, object masks, and grasp mode priors. Park et al. [17] introduced Learned Geometric Feasibility (LGF), a voxel-based model that predicts grasp feasibility from local occupancy and supports logic-geometric planning.

Although these feasibility-aware approaches helped filter out invalid candidates and reduce planning time, they assumed a small predefined grasp set and had limited scalability. Our method addresses this limitation by enabling grasp prediction over a grasp set automatically annotated using sampling methods. It learns a differentiable cost landscape for task-aware scoring and grasp selection in the object's local frame, and is generalizable to denser grasp sets than those used during training.

B. Grasp Selection

Unlike grasp generation from scratch, grasp selection aims to identify feasible grasps from a predefined candidate set, while considering object and robotic constraints. Previous studies mainly focus on grasp selection for the picking phase. Herzog et al. [18] and Chen et al. [19] retrieved grasps by comparing object geometries with known instances. Gualtier et al. [20] and Merwe et al. [21] used learning to extract grasp-relevant features from object geometry. More recently, Qian et al. [22] used a large language model to select grasps from a predefined candidate set in a semantically informed context.

In tasks that involve both picking and placing, grasp selection becomes a core component of the overall planning pipeline. He et al. [23] selected grasp by maximizing placement affordance, ensuring feasibility under both the initial and goal object poses. Xu et al. [24] proposed a reinforcement learning framework that encodes both initial and goal states and learns grasp selection policies in an end-to-end manner. For more complex regrasping problems, Wan et al. [25] and Cheng et al. [26] represented grasp-object configurations as nodes in a graph and performed search to identify feasible regrasp sequences. Xu et al. [27] extended this idea to hierarchical planning by jointly optimizing grasp pairs and intermediate object poses, guided by a learned cost estimator.

While these methods typically model grasp selection as a joint optimization over initial and goal poses, our approach adopts a different perspective. We observed that (i) the majority of candidate grasps are infeasible across pick-and-place pairs and (ii) learning over the full joint space requires a large amount of data. We therefore treated the placing process as reversed picking and formulated the joint selection problem into two independent per-pose evaluations. The formulation enabled efficient composition of feasible grasps and thus accelerated the identification of shared grasps.

III. Modeling Feasible Grasps Using EBM

Mathematically, given the initial and goal object poses $\boldsymbol{T}_{\text{init}}, \boldsymbol{T}_{\text{goal}} \in SE(3)$, and a set of candidate grasp $\boldsymbol{\mathcal{G}}_0 = \{(\boldsymbol{g}, w)\}$, where $\boldsymbol{g} \in SE(3)$ represents the gripper's pose in the object's canonical coordinate frame $\boldsymbol{T}_0 = \boldsymbol{I}_{4\times4}$, and $w \in \mathbb{R}$ denotes the corresponding normalized gripper width (i.e., finger opening), we aim to predict the subset $\boldsymbol{\mathcal{G}}_{\text{shared}} \subseteq \boldsymbol{\mathcal{G}}_0$ that remains feasible when transformed under both the initial and goal object poses. In detail, a grasp $(\boldsymbol{g}, w) \in \boldsymbol{\mathcal{G}}_0$ is considered a *shared grasp* if both the transformed grasp poses $(\boldsymbol{T}_{\text{init}} \cdot \boldsymbol{g}, w)$ and $(\boldsymbol{T}_{\text{goal}} \cdot \boldsymbol{g}, w)$ satisfy the IK and collision constraints. To predict the shared grasps, our method adopts a two-stage learning approach. We first train a model to predict, for a given object pose $\boldsymbol{T}$, whether the transformed grasp pose $(\boldsymbol{T} \cdot \boldsymbol{g}, w)$ satisfies the feasibility constraints. This per-pose feasibility prediction is then extended to shared grasp prediction by evaluating grasp candidates under both $\boldsymbol{T}_{\text{init}}$ and $\boldsymbol{T}_{\text{goal}}$. In the remainder of this section, we focus on modeling the feasible grasp prediction. In the next section, we will discuss combining the predictions under different object poses to construct the shared grasp set.

We use EBM to model the feasible grasp. An EBM defines an energy function over object and grasp pairs and assigns lower energy to more feasible ones. Unlike traditional discriminative models that directly predict a mask vector, an EBM fits the joint probability of object and grasp using

$$p_\phi(x) = \frac{\exp(-E_\phi(x)/t)}{Z_\phi}, \tag{1}$$

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

where x denotes the input variables. $E_\phi(x)$ is a learnable energy function parameterized by ϕ, t is the Boltzmann temperature constant and constant Z_ϕ is defined as

$$Z_\phi = \int \exp(-E_\phi(x)/t)dx. \tag{2}$$

It serves as a normalization constant to ensure that $p_\phi(x)$ integrates to one. In practice, Z_ϕ is often intractable to compute when the output space is continuous or high-dimensional. However, in our setting, both the input and output spaces are discretized, which allows this integral to be approximated by a summation over a finite set of candidates. In the context of grasp planning, to facilitate training, here we set $x = [\boldsymbol{T}, \boldsymbol{g}, w]$, The goal of the EBM is to learn an energy function $E_{\phi_f}(\boldsymbol{T}, \boldsymbol{g}, w)$ that assigns low energy to feasible grasp – those that satisfy IK and collision constraints – under the given object pose. The energy function $E_{\phi_f}(\boldsymbol{T}, \boldsymbol{g}, w)$ is implemented as a neural network that takes as input the concatenation of a pose encoding of $\boldsymbol{T}$, a grasp pose $\boldsymbol{g}$, and gripper width w. It is trained to output a scalar energy value.

To train the energy network, we adopt a Negative Log-Likelihood (NLL) loss function that encourages the model to assign low energy to feasible samples

$$\mathcal{L}_{\text{nll}} = \mathbb{E}_{\boldsymbol{T}_j}\left[\mathbb{E}_{(\boldsymbol{g},w)\text{ feasible for }\boldsymbol{T}_j}\left[\frac{E_{\phi_f}(\boldsymbol{T}_j, \boldsymbol{g}, w)}{t}\right]\right] + \log Z_{\phi_f}, \tag{3}$$

where the partition function Z_{ϕ_f} is approximated as

$$Z_{\phi_f} = \sum_{\boldsymbol{T}_j}\sum_{(\boldsymbol{g},w)} \exp\left(-\frac{E_{\phi_f}(\boldsymbol{T}_j, \boldsymbol{g}, w)}{t}\right). \tag{4}$$

The $\mathbb{E}$ in equation (3) represents expectational computations. They are taken over all object poses $\boldsymbol{T}_j$ in the training set, and the set of grasp candidates $(\boldsymbol{g}, w)$ that are feasible for each $\boldsymbol{T}_j$. The summations in equation (4) are taken over all object–grasp pairs encountered during training, regardless of feasibility.

Meanwhile, we use a contrastive energy loss defined as

$$\mathcal{L}_{\text{con}} = \mathcal{L}_+ - \mathcal{L}_-, \tag{5}$$

where

$$\begin{cases} \mathcal{L}_+ = \mathbb{E}_{\boldsymbol{T}_j}\left[\mathbb{E}_{(\boldsymbol{g},w)\text{ feasible for }\boldsymbol{T}_j}\left[\frac{E_{\phi_f}(\boldsymbol{T}_j, \boldsymbol{g}, w)}{t}\right]\right] \\ \mathcal{L}_- = \mathbb{E}_{\boldsymbol{T}_j}\left[\mathbb{E}_{(\boldsymbol{g},w)\text{ infeasible for }\boldsymbol{T}_j}\left[\frac{E_{\phi_f}(\boldsymbol{T}_j, \boldsymbol{g}, w)}{t}\right]\right] \end{cases}, \tag{6}$$

to assign lower energy to feasible grasps and higher energy to infeasible ones.

In addition, we apply an energy regulation term to prevent divergence

$$\mathcal{L}_{\text{reg}} = \mathcal{L}_+^2 + \mathcal{L}_-^2, \tag{7}$$

where

$$\begin{cases} \mathcal{L}_+^2 = \mathbb{E}_{\boldsymbol{T}_j}\left[\mathbb{E}_{(\boldsymbol{g},w)\text{ feasible for }\boldsymbol{T}_j}\left[\left(\frac{E_{\phi_f}(\boldsymbol{T}_j, \boldsymbol{g}, w)}{t}\right)^2\right]\right] \\ \mathcal{L}_-^2 = \mathbb{E}_{\boldsymbol{T}_j}\left[\mathbb{E}_{(\boldsymbol{g},w)\text{ infeasible for }\boldsymbol{T}_j}\left[\left(\frac{E_{\phi_f}(\boldsymbol{T}_j, \boldsymbol{g}, w)}{t}\right)^2\right]\right] \end{cases}. \tag{8}$$

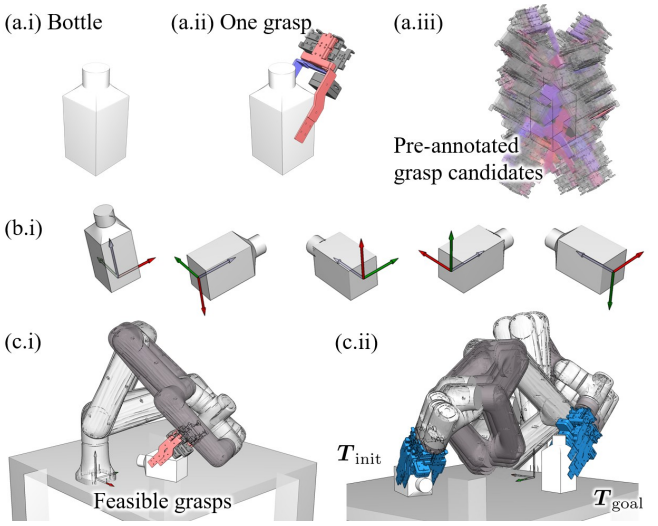

Fig. 2: Grasp dataset collection for the bottle. (a.i) Bottle object. (a.ii) Collision-free grasp candidate generated by sampling the parallel mesh facets. (a.iii) Grasp candidate set $\mathcal{G}_0$. (b.i) Stable placements of the object before considering yaw rotations. (c.i) Feasible grasps. (c.ii) Shared grasps.

The final loss is a weighted sum

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{nll}} + \mathcal{L}_{\text{con}} + \alpha\mathcal{L}_{\text{reg}}, \tag{9}$$

where α is the proportion constant of the regulation term.

During inference, we follow standard binary classification principles to decide whether the given input is feasible. Specifically, the learned energy model E_{ϕ_f} assigns a scalar energy score to each input, and we classify it as feasible if the score is below a threshold h_f determined using the validation set[FT1].

For training data collection, we follow the method proposed by Wan et al. [2] to generate grasp candidates $\{(\boldsymbol{g}, w)\}$ and object poses $\{\boldsymbol{T}_j\}$. The overall pipeline is illustrated in Fig. 2(a.i–a.iii). Given a 3D object mesh, we uniformly sample surface points and extract antipodal contact pairs as initial grasp candidates. For each pair, we align the gripper to the contact points and generate an initial grasp pose by transforming it to the contact frame with a random rotation around the antipodal axis (Fig. 2(a.ii)). We then further rotate the gripper around the contact normal and perform collision checking. Grasps that are collision-free and satisfy the force-closure condition are retained (Fig. 2(a.iii)). This procedure yields N candidate grasps per object.

To generate object poses, we assume a planar surface and compute the stable placements of the object (Fig. 2(b.i)). The placements are then diversified by randomly sampling planar positions and in-plane yaw angles, which helps effectively cover the SE(2) space. In total, we produce M object poses.

These object-grasp pairs are used to train the EBM with the loss function described previously. Fig. 2(c.i) shows an example object pose and its corresponding feasible grasps. Fig. 2(c.ii) illustrates the resulting shared grasp set under a given initial and goal pose pair.

[FT1] We evaluate the F1 score on a validation set for a range of threshold values, and select the one that yields the highest F1 score as h_f.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

IV. Shared Grasp Prediction

This section introduces the proposed method for predicting the shared grasp set. The prediction is formulated as a joint probability estimation problem using compositional EBMs, and the goal is to identify grasps that are simultaneously feasible under both the initial and goal object configurations.

In particular, we define the shared grasp probability as a joint distribution over the initial pose $\boldsymbol{T}_{\text{init}}$, goal pose $\boldsymbol{T}_{\text{goal}}$, and a grasp candidate $(\boldsymbol{g}, w)$. Under the compositional EBM framework, this joint distribution is approximated as a product of independent terms:

$$p(x_1, x_2, ..., x_n) = \prod\nolimits_i p_\phi(x_i) \propto \exp\left(-\sum\nolimits_i E_\phi(x_i)\right), \tag{10}$$

which, in our context, gives:

$$p(\boldsymbol{T}_{\text{init}}, \boldsymbol{T}_{\text{goal}}, \boldsymbol{g}, w) = p(\boldsymbol{T}_{\text{init}}, \boldsymbol{g}, w) \cdot p(\boldsymbol{T}_{\text{goal}}, \boldsymbol{g}, w). \tag{11}$$

Following the Boltzmann distribution equation (1), this expression leads to an additive energy representation:

$$p(\boldsymbol{T}_{\text{init}}, \boldsymbol{T}_{\text{goal}}, \boldsymbol{g}, w) \propto \exp\left(-\left[E_{\phi_f}(\boldsymbol{T}_{\text{init}}, \boldsymbol{g}, w) + E_{\phi_f}(\boldsymbol{T}_{\text{goal}}, \boldsymbol{g}, w)\right]\right). \tag{12}$$

The addition implies that given a candidate $(\boldsymbol{g}, w)$, we can compute its total energy as the sum of its energies under both object poses. Accordingly, the prediction of the shared grasp set is formulated as selecting grasp candidates whose joint energy is lower than a predefined threshold h_s:

$$\boldsymbol{\mathcal{G}}_{\text{shared}} = \left\{(\boldsymbol{g}, w) \mid E_{\phi_f}(\boldsymbol{T}_{\text{init}}, \boldsymbol{g}, w) + E_{\phi_f}(\boldsymbol{T}_{\text{goal}}, \boldsymbol{g}, w) < h_s\right\}. \tag{13}$$

It is important to note that h_s differs from the feasible grasp threshold h_f used in the previous section. In the joint framework, we evaluate the summed energy over both poses and apply a single threshold h_s to determine shared grasps without per-pose feasibility prediction. The selection of h_s also follows maximizing the F1 score over a validation set[FT2].

During inference (Fig. 3), all input tuples $[\boldsymbol{T}_{\text{init}}, \boldsymbol{g}, w]$ and $[\boldsymbol{T}_{\text{goal}}, \boldsymbol{g}, w]$ for N grasp candidates are packed into a single tensor and passed through the energy model in one forward pass for efficiency. The resulting $2N$ energy values are classified using the threshold h_s, yielding a binary mask $\hat{y}$ used to extract the final shared grasp set: $\boldsymbol{\mathcal{G}}_{\text{shared}} = \boldsymbol{\mathcal{G}}_0[\hat{y} = 1]$.

It is worth noting that the above method represents one possible implementation of shared grasp prediction. In addition to this joint estimation approach, we also consider two variations: a direct prediction method that models shared grasp energy in a single step, and a logical conjunction method that classifies shared grasps based on per-pose feasibility. These variations are described in detail in Appendix I and are quantitatively compared in the experiments.

[FT2]However, since shared grasp prediction involves evaluating grasp feasibility under both initial and goal poses, the computation of F1 must be based on grasp sets that are truly shared across poses. The dataset used to train the EBM model only involved feasible grasp datasets. It lacked information about shared grasps. For this reason, we additionally synthesize a ground-truth shared grasp dataset by randomly sampling object pose pairs $(\boldsymbol{T}_{\text{init}}, \boldsymbol{T}_{\text{goal}})$, analytically evaluating their corresponding feasible grasp sets based on $\boldsymbol{G}_0$, and then computing the intersection of the feasible sets. The resulting dataset, $\{\boldsymbol{T}_{\text{init}}, \boldsymbol{T}_{\text{goal}}, \boldsymbol{g}, w\}$, provides supervision for evaluating classification performance under joint energy and is used to determine h_s.

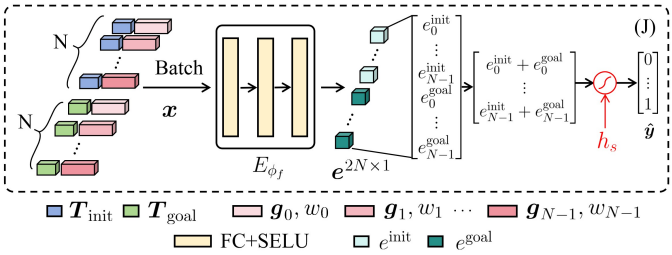

Fig. 3: Proposed method uses an EBM E_{ϕ_f} to independently model the energy values of pre-annotated grasp candidates under each object pose and sum up corresponding energy items to jointly estimate the energy value of the shared grasp. Then, the method employs a binary classification strategy to determine whether each grasp qualifies as a shared one.

V. Experiments and Analysis

We use a 6-DOF Dobot Nova2 robotic arm equipped with a two-finger gripper for the experiments. The training data are collected in the simulation environment illustrated in Fig. 1. The sampling space is constrained to $x \in [-0.45, 0.45]$ m, $y \in [0.1, 0.6]$ m, and $\theta \in [0, 2\pi]$ rad. The sampling resolution is set to 0.001 m for position and 0.01 rad for orientation. The computational setup used for the experiments consisted of an Intel 13th Gen Core i9-13900KF processor with 64 GB of RAM and an NVIDIA RTX 4090 GPU. Our EBM is implemented using a three-layer fully connected neural network with SELU activation functions. Training is performed with a batch size of 1024, a learning rate of 1×10^{-3}, a temperature constant of 0.5, and a regularization coefficient $\alpha = 0.2$.

The experiments are divided into two parts. First, we evaluate the overall performance of the proposed joint estimation method (J method). We begin by comparing it with analytical baselines to validate the effectiveness of the EBM-based prediction pipeline. We then compare the J method with two alternative formulations, the direct prediction method (D method) and the logical conjunction method (L method), to examine how different strategies for leveraging EBM outputs affect prediction accuracy and data efficiency under varying training data ratios. Second, we analyze the generalization ability of the J, D, and L methods by testing their performance on unseen grasps and unseen objects.

A. Performance

1) Comparison with Analytical Methods: We first compare the J method with analytical baselines to evaluate the effectiveness of the EBM-based shared grasp prediction. The experiments are conducted using the bottle-shaped object shown in Fig. 2. To investigate how the size of the grasp candidate set $\boldsymbol{\mathcal{G}}_0$ affects prediction performance, we pre-generate three sets containing 57, 109, and 352 grasp candidates, respectively. For training the J method, we randomize the object's position and orientation on the table and transform grasp candidates to these poses to generate feasible grasps. For each set of candidate grasps, we collect 75k feasible grasps, among which 50k are used for training, 15k for testing, and 10k for validation (i.e., for synthesizing the ground-truth shared grasp dataset

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

and determining h_s). To evaluate the grasp selection strategy, we compare two J variations: (a) J^{R}, which randomly selects a grasp from the predicted positive set, and (b) J^{O}, which selects the grasp with the lowest predicted energy among the positive set. As for the analytical baselines, we consider three implementations: (i) R method, which directly samples a grasp uniformly at random from the candidate set without any IK or collision filtering, and (ii) A method, which filters candidates by computing the IK solutions and checking for collisions at both the initial and goal object poses, and then selects a grasp randomly from the intersection of feasible candidates. (iii) H method, which is a heuristic approach that prioritizes grasp candidates based on their proximity to the robot's initial TCP at both the pick and place poses. For each candidate, we compute a heuristic score as the sum of the normalized distances at the two poses. Candidates are ranked in ascending order of this score, and those with lower values are preferentially selected for motion planning.

Table I presents the results. The columns are grouped into four major sections, each corresponding to a different size of $\mathcal{G}_0$. The $\overline{S}$ row reports the success rate of finding a collision-free and IK-feasible shared grasp over 1000 trials for each method. The $\overline{t}$ row shows the average time required to find a successful solution. We can see that the proposed methods are faster than all baselines and achieve consistently high success rates across different grasp candidate numbers.

TABLE I: Comparison with analytical methods

	57 grasp candidates					109 grasp candidates					352 grasp candidates				
	R	A	H	J^{R}	J^{O}	R	A	H	J^{R}	J^{O}	R	A	H	J^{R}	J^{O}
$\overline{S_g}$	9.5	100	30.5	86.6	92.7	7.0	100	32.5	87.4	88.4	6.2	100	33.0	81.1	89.9
$\overline{t_g}$	–	5.7	2.5	1.7		–	11.0	4.5	3.1		–	34.1	15.8	9.2	

Note 1 R – Random selection from grasp candidates, A – Random selection from the intersection of grasps that are IK-feasible and collision-free at both initial and goal poses, H – Heuristic sample (colosed grasp pose), J^{R} – Random selection from the result of J, J^{O} – Selection of the minimum-energy grasp from the result of J.
Note 2 The best value in each comparison block is highlighted in lime.

Table II reports the success rates and planning time when using parallel RRT-C (with up to 30 parallel processes due to system limits) to generate the motion based on the selected grasps. The top 8, 16, and 24 grasps of respective methods from the dataset of 57 candidates were evaluated. The results show that the grasps selected by J lead to less planning time. We believe this is because J consistently selects low-energy grasps. Such grasps tend to lie deeper inside the collision-free and IK-feasible region, making motion planning easier.

TABLE II: Effect on motion planning

	8 / 57		16 / 57		24 / 57	
	succ. rate	time (s)	succ. rate	time (s)	succ. rate	time (s)
R + pRRT-C	37.0%	9.25	48.0%	9.10	60.0%	9.59
A + pRRT-C	78.0%	9.12	78.0%	9.23	78.0%	9.40
J + pRRT-C	77.0%	9.08	77.0%	9.12	78.0%	9.15
H + pRRT-C	72.0%	9.23	75.0%	9.69	76.0%	9.71

2) Comparison with Other Prediction Variations: We next compare the proposed J method with the D and L alternative formulations to investigate how different strategies for leveraging the EBM outputs influence grasp prediction accuracy and data efficiency. Detailed implementation of both methods is provided in the Appendix. Similar to the previous experiments, we use the bottle as the target object. To ensure a fair comparison, we fix the size of the grasp candidate set $\mathcal{G}_0$ to 57 pre-generated grasp candidates. For J and L, we randomly sampled object poses and collected 280k feasible grasps. For D, we randomly sampled object pose pairs and collected 280k shared grasps. Each dataset was split into 200k for training, 50k for testing, and 30k for validation. To better understand the data efficiency and precision of each method, we further subdivide the 200k training data into varying proportions and evaluate the performance under different training data ratios.

Table III shows the results. We can see from the table that the J method, even when trained on only 50% of the 200k training data, outperforms both the D and L methods trained with the full 100% dataset. This demonstrates the high data efficiency of the J method. We also observe that the D method exhibits comparable performance to the L method. However, the D method inherently requires twice the data collection time, as it needs to generate shared grasp labels over pose pairs, making it the least data-efficient among the three. The L method essentially performs two independent thresholding operations, one applied to the classifier at the initial pose and the other to the classifier at the goal pose. The separation helps filter out more near-threshold false negatives, which likely contributes to the method's higher recall compared to the J method. However, it increases strictness and leads to a reduction in precision. The "F" row of the table shows the performance of a network trained solely to predict feasible grasps for reference. From the results, we conclude that the J method offers a relatively balanced performance across recall and precision, which is why we mainly presented it in the main text. Nevertheless, we do not consider the D and L methods to be inferior. In the following generalization experiments, we include all three methods for a comprehensive comparison.

TABLE III: Comparison of other prediction methods.

	100%			50%			15%			5%		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
J	94.0	95.3	94.6	94.2	91.3	93.0	90.0	81.9	85.7	85.9	77.8	81.6
D	92.9	95.2	94.1	91.2	94.4	92.8	78.9	81.5	80.2	70.1	80.4	74.9
L	90.7	97.3	93.8	89.7	96.5	93.0	78.5	93.2	85.2	77.4	88.5	82.6
F	98.2	98.7	98.4	98.0	98.4	98.2	94.5	96.6	95.5	92.1	94.0	93.0

Note F – Performance of the EBM model trained only for feasible grasp prediction, without incorporating shared grasp considerations.

3) Ablation study for grasp stability: We also compared the real-world success rate and pick-and-place accuracy of the bottle, bunny, power drill, and mug from the YCB object set [28]. For each object, we prepared 180 grasp candidates and based on them, we collected 150k data per object, including feasibility annotations and shared grasp labels. Fig. 4 illustrates the objects and exemplifies placements, grasp

candidates, feasible grasps, and shared grasps.

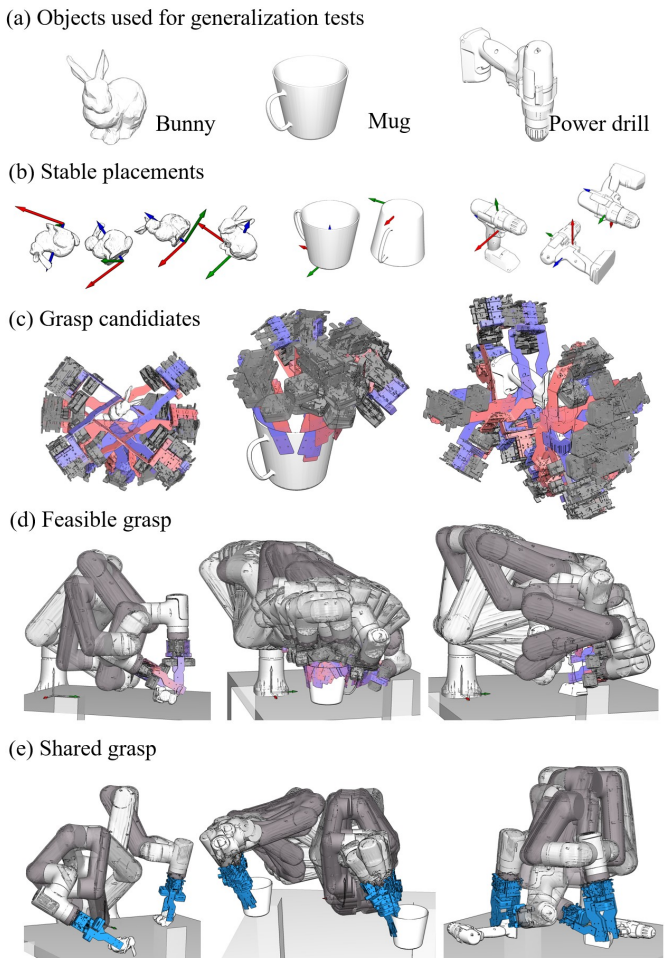

Fig. 4: Dataset collection for Bunny, Mug, and Power drill.

Besides the J^R and J^O variations, we further compared the J^C method, which ranks and selects predicted grasps based on the distance between the grasp center and the object's Center of Mass (CoM). The method helps ensure better grasp stability. From the results shown in Table IV we can see that all J variations are generally effective. While considering CoM distances provides some benefit, the improvement is not substantial for our objects. It is a choice to be flexibly adjusted according to application requirements. We can also see from the results that the Bunny had the worst performance. We believe this is because the current gripper had limited stability and could not adapt to the Bunny's rounded contour. Extending the framework to more advanced stability models while considering gripper designs could be an interesting future direction. The last two columns of the table report the mean error and deviation measured using FoundationPose [29] for successfully completed tasks.

B. Generalization

1) Unseen Grasps: We were interested in examining whether the EBM-based methods can generalize across the grasp space of the bottle and identify unseen feasible grasp poses. To this end, we trained models using grasp candidate

TABLE IV: Success rates and errors for four objects

	Pick ($J^R \cdot J^O \cdot J^C$)	Place ($J^R \cdot J^O \cdot J^C$)	Δpos (mm)	Δrot (°)
Bt	7/10 · 9/10 · 10/10	7/10 · 9/10 · 10/10	2.2±1.2	1.2±0.4
Bn	2/10 · 1/10 · 2/10	2/10 · 1/10 · 2/10	9.3±3.1	7.8±4.5
Pd	9/10 · 10/10 · 10/10	9/10 · 10/10 · 10/10	5.1±2.1	2.8±2.2
Mg	10/10 · 10/10 · 10/10	10/10 · 10/10 · 10/10	4.6±2.1	4.1±2.4

Note 1 Bt – Bottle, Bn – Bunny, Pd – Power drill, Mg – Mug.
Note 2 J^C – Selection based distances between CoMs and grasping centers.
Note 3 Success rates are measured over 10 times of executions.
Note 4 Δpos, Δrot – Position and rotation errors at placement.

sets of size 57, 83, 109, and 352, and evaluated them on a separate set comprising 922 candidates[FT3]. The evaluation results are summarized in Table V[FT4]. We can see that increasing the number of training grasp candidates consistently improves generalization to unseen grasp poses. Under the same training conditions, the J method achieves the best performance. In contrast, the D method requires a larger amount of data to perform well, while the L method tends to be overly conservative due to its strict decision criterion.

TABLE V: Generality to unseen grasp poses.

	Model - 1			Model - 2			Model - 3			Model - 4		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
J	72.7	62.7	67.3	76.8	45.5	57.2	81.1	71.2	75.8	89.2	86.3	87.8
D	65.1	45.5	53.6	67.9	39.5	50.0	72.8	54.6	62.4	76.2	66.1	70.7
L	65.5	62.6	64.0	74.8	45.1	56.3	76.3	67.7	71.7	81.0	74.2	77.4
F	97.1	98.3	97.7	97.9	98.2	98.1	97.7	98.3	98.1	97.8	98.3	98.2

Note 1 Model 1 $\sim$ 4 trained on 57, 83, 109, 352 grasp candidates;
Note 2 The best values for shared grasp prediction are highlighted in lime.

2) Varying Heights: We evaluated performance under varying support surface heights. Our objects used include the bottle plus the milk box, bowl, pitcher, bunny, and power drill from the YCB set. For each object, we collected about 20k feasible grasps by varying the surface height from 0 mm to 200 mm in 50 mm increments, and trained a separate model. Fig. 5 shows examples of data collection for the bunny and milk box. For testing, we used intermediate heights ranging from 25 mm to 225 mm in 50 mm increments, collecting about 1k feasible grasps per object. The results in Table VI indicate that the proposed method maintains good generalization performance across different heights. Although the training and testing heights were not identical, the models still achieved consistently high P, R, and F1 scores.

3) Unseen Objects: We also evaluated the generalization capability of the three methods under variations in object geometry. The same datasets used for grasp stability studies were used. To assess cross-object generalization, we trained the models using different combinations of object datasets

[FT3]For training, the J and L methods each used 75k feasible grasp samples per candidate set. The D method was trained on 75k shared grasp samples. All datasets were split into 50k for training, 15k for testing, and 10k for validation. The evaluation on the 922-grasp set was conducted using 5k samples.

[FT4]For Model-4 of the J method, we increased the evaluation set size to 10,032 candidates and observed no significant performance degradation. The resulting precision, recall, and F1 score were 90.2, 84.6, and 87.3, respectively, which are comparable to the results obtained with 922 candidates. This further confirms that our method generalizes well to denser grasp sets.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

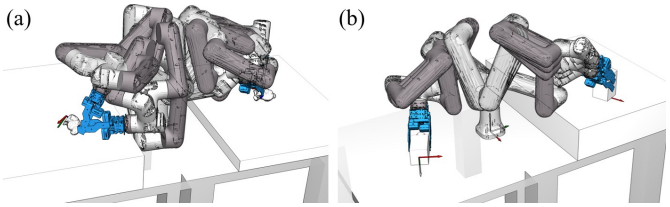

Fig. 5: Collecting data of different surface heights. The two sub-figures show two objects: (a) Bunny. (b) Milk box.

TABLE VI: Generalization on surfaces of different heights

	P	R	F1		P	R	F1		P	R	F1
Bt	85.0	86.3	85.6	Mk	91.2	95.8	93.4	Bw	95.3	93.6	94.4
Pt	83.6	78.4	80.8	Bn	82.4	83.4	82.9	Pd	71.7	87.6	78.9

Note Mk – Milk box; Bw – Bowl; Pt – Pitcher. See Table IV for others.

and tested their performance on the held-out (unseen) ones. Table VII summarizes the results. The "Dataset" column lists the object combinations used for training. The "Bunny," "Mug," and "Power drill" columns report scores of the trained models evaluated on each object. Gray text denotes seen (in-distribution) objects. Black text denotes unseen ones.

TABLE VII: Generality to unseen objects.

	Dataset	Bunny			Mug			Power drill		
		P	R	F1	P	R	F1	P	R	F1
J	Bt	1.6	100	3.2	53.3	62.1	57.3	23.6	91.9	37.6
	Bt+Bn	88.2	97.9	92.8	63.0	50.3	55.9	26.9	16.8	20.7
	Bt+Mg	1.7	80.5	3.4	80.0	97.7	86.7	20.8	87.7	33.7
	Bt+Pd	1.6	98.8	3.1	32.9	98.8	49.4	75.7	95.8	84.6
	Bt+Bn+Mg	80.6	98.8	88.8	76.6	96.6	85.5	29.6	47.4	36.4
	Bt+Mg+Pd	1.8	89.1	3.4	74.8	97.5	84.7	73.9	96.2	83.6
	Bt+Bn+Pd	85.8	97.2	91.1	47.1	54.5	50.5	74.2	95.1	83.4
D	Bt	4.2	0.5	0.1	14.7	6.7	9.2	21.5	69.0	32.8
	Bt+Bn	82.1	55.2	66.0	75.0	4.2	7.9	0.0	0.0	0.0
	Bt+Mg	0.0	0.0	0.0	78.6	85.7	82.0	0.0	0.0	0.0
	Bt+Pd	6.5	2.1	3.2	47.4	0.9	1.7	74.2	75.8	75.0
	Bt+Bn+Mg	80.6	42.0	55.2	76.0	82.8	79.3	0.0	0.0	0.0
	Bt+Mg+Pd	0.0	0.0	0.0	76.1	81.9	78.9	71.6	70.1	70.8
	Bt+Bn+Pd	85.3	52.7	65.1	36.9	0.8	1.6	71.3	73.9	72.5
L	Bt	6.2	7.6	6.8	84.4	39.6	53.9	45.5	17.7	25.5
	Bt+Bn	91.6	81.6	86.3	88.6	17.4	29.1	0.0	0.0	0.0
	Bt+Mg	5.6	0.6	1.2	80.6	92.3	86.0	0.0	0.0	0.0
	Bt+Pd	2.1	0.3	0.5	87.8	10.1	18.1	78.8	88.8	83.5
	Bt+Bn+Mg	88.9	72.0	79.5	79.6	91.7	85.2	0.0	0.0	0.0
	Bt+Mg+Pd	0.8	0.3	0.5	79.2	90.4	84.4	77.5	85.5	81.3
	Bt+Bn+Pd	90.3	76.6	82.9	72.7	6.7	12.3	78.0	87.6	82.6

Note 1 Bt – Bottle; Bn – Bunny; Mg – Mug; Pd – Power drill;
Note 2 Gray: Results of seen objects; Black: Results of unseen objects.

The results show that all methods generalize reasonably well to the mug object. We attribute this to its geometric similarity to the bottle used in training, as both share comparable feasible grasp distributions. In contrast, generalization to less similar shapes (e.g., bunny and drill) is more challenging.

Among the three methods, the D method performs the worst on unseen objects, suggesting that directly predicting shared grasps lacks robustness under limited training data. The L method shows strong performance on in-distribution data but suffers degraded generalization to the mug when trained jointly with dissimilar object data. This is likely due to interference across shape domains, which affects threshold tuning. The J method, despite slightly lower precision, demonstrates better overall generalization across different objects, making it more suitable for scaling to diverse object categories.

VI. Conclusions and Future Work

We proposed an EBM-based method to model feasible grasps and predict shared grasps by composing the learned energies of two feasible picks. Compared to analytical baselines, our approach has higher efficiency and reliable accuracy. It can generalize to unseen grasps of objects with similar shapes. Conceptually, the method is equivalent to using reachability learning for shared grasp prediction, which naturally enables partial cross-object generalization. Meanwhile, the method provides tangible efficiency gains for practical pick-and-place planning.

While our current design relies on predefined grasp candidates to ensure feasibility and robustness, this comes at the cost of adaptability. A promising future direction is to integrate adaptive candidate generation to enable greater flexibility while preserving the stability and efficiency of the current framework. Embedding stability directly into EBM is also an interesting solution. However, this would significantly increase model complexity, and need intensive exploration. Moreover, we are interested in extending the prediction model to include object-shape information and collision constraints with surrounding obstacles in future work.

References

[1] J. E. King, M. Klingensmith, C. M. Dellin, M. R. Dogar, P. Velagapudi, N. S. Pollard, and S. S. Srinivasa, "Pregrasp manipulation as trajectory optimization." in *RSS*, 2013.

[2] W. Wan, M. T. Mason, R. Fukui, and Y. Kuniyoshi, "Improving regrasp algorithms to analyze the utility of work surfaces in a workcell," in *ICRA*, 2015, pp. 4326–4333.

[3] P. Xu, Z. Chen, J. Wang, and M. Q.-H. Meng, "Learning to predict diverse stable placements for extrinsic manipulation on a support plane," *IEEE Trans. Cogn. Dev. Syst.*, vol. 16, no. 3, pp. 1095–1107, 2023.

[4] M. D. Shanthi and T. Hermans, "Pick and place planning is better than pick planning then place planning," *IEEE Robot. Autom. Lett.*, vol. 9, no. 3, pp. 2790–2797, 2024.

[5] E. Maranci, S. D'Avella, P. Tripicchio, C. Avizzano *et al.*, "Enabling grasp synthesis approaches to task-oriented grasping considering the end-state comfort and confidence effects," *IEEE Robot. Autom. Lett.*, vol. 9, no. 6, pp. 5695–5702, 2024.

[6] B. H. Leebron, K. Ren, Y. Chen, and K. Hang, "B4p: Simultaneous grasp and motion planning for object placement via parallelized bidirectional forests and path repair," *arXiv:2504.04598*, 2025.

[7] H. Chen, W. Wan, K. Koyama, and K. Harada, "Planning to build block structures with unstable intermediate states using two manipulators," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 4, pp. 3777–3793, 2022.

[8] M. Costanzo, S. Stelter, C. Natale, S. Pirozzi, G. Bartels, A. Maldonado, and M. Beetz, "Manipulation planning and control for shelf replenishment," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 1595–1601, 2020.

[9] B. Kim, Z. Wang, L. P. Kaelbling, and T. Lozano-Pérez, "Learning to guide task and motion planning using score-space representation," *Int. J. Robot. Res.*, vol. 38, no. 7, pp. 793–812, 2019.

[10] Z. Yang, C. Garrett, T. Lozano-Perez, L. Kaelbling, and D. Fox, "Sequence-based plan feasibility prediction for efficient task and motion planning," in *RSS*, 2023.

[11] M. Khodeir, B. Agro, and F. Shkurti, "Learning to search in task and motion planning with streams," *IEEE Robot. Autom. Lett.*, vol. 8, no. 4, pp. 1983–1990, 2023.

[12] A. M. Wells, N. T. Dantam, A. Shrivastava, and L. E. Kavraki, "Learning feasibility for task and motion planning in tabletop environments," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1255–1262, 2019.

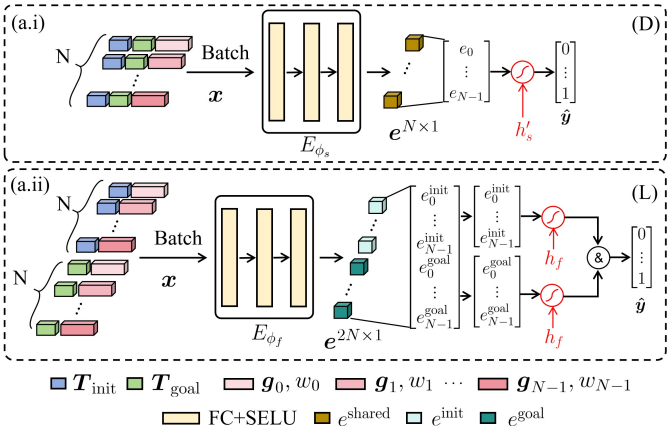

Fig. 6: (a.i) Direct Prediction estimates the energy value of shared grasps using a unified EBM model. (a.ii) Logical Conjunction uses an EBM to estimate the feasibility of each object pose and grasp candidate pair, and then employs a Logical AND for predicting the shared ones.

[13] D. Driess, O. Oguz, J.-S. Ha, and M. Toussaint, "Deep visual heuristics: Learning feasibility of mixed-integer programs for manipulation planning," in *ICRA*, 2020, pp. 9563–9569.

[14] L. Xu, T. Ren, G. Chalvatzaki, and J. Peters, "Accelerating integrated task and motion planning with neural feasibility checking," *arXiv:2203.10568*, 2022.

[15] S. Ait Bouhsain, R. Alami, and T. Simeon, "Learning to predict action feasibility for task and motion planning in 3d environments," in *ICRA*, 2023, pp. 3736–3742.

[16] ——, "Simultaneous action and grasp feasibility prediction for task and motion planning through multi-task learning," in *IROS*, 2023, pp. 2042–2048.

[17] S. Park, H. C. Kim, J. Baek, and J. Park, "Scalable learned geometric feasibility for cooperative grasp and motion planning," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 11 545–11 552, 2022.

[18] A. Herzog, P. Pastor, M. Kalakrishnan, L. Righetti, T. Asfour, and S. Schaal, "Template-based learning of grasp selection," in *ICRA*, 2012, pp. 2379–2384.

[19] H. Chen, T. Kiyokawa, W. Wan, and K. Harada, "Category-association based similarity matching for novel object pick-and-place task," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 2961–2968, 2022.

[20] M. Gualtieri, A. Ten Pas, K. Saenko, and R. Platt, "High precision grasp pose detection in dense clutter," in *IROS*, 2016, pp. 598–605.

[21] M. Van der Merwe, Q. Lu, B. Sundaralingam, M. Matak, and T. Hermans, "Learning continuous 3d reconstructions for geometrically aware grasping," in *ICRA*, 2020, pp. 11 516–11 522.

[22] Y. Qian, X. Zhu, O. Biza, S. Jiang, L. Zhao, H. Huang, Y. Qi, and R. Platt, "Thinkgrasp: A vision-language system for strategic part grasping in clutter," *arXiv:2407.11298*, 2024.

[23] Z. He, N. Chavan-Dafle, J. Huh, S. Song, and V. Isler, "Pick2place: Task-aware 6dof grasp estimation via object-centric perspective affordance," in *ICRA*, 2023, pp. 7996–8002.

[24] K. Xu, Z. Zhou, J. Wu, H. Lu, R. Xiong, and Y. Wang, "Grasp, see, and place: Efficient unknown object rearrangement with policy structure prior," *IEEE Trans. Robot.*, vol. 41, pp. 464–483, 2025.

[25] W. Wan, K. Harada, and F. Kanehiro, "Preparatory manipulation planning using automatically determined single and dual arm," *IEEE Trans. Ind. Inform.*, vol. 16, no. 1, pp. 442–453, 2019.

[26] S. Cheng, K. Mo, and L. Shao, "Learning to regrasp by learning to place," in *CoRL*, 2022, pp. 277–286.

[27] K. Xu, H. Yu, R. Huang, D. Guo, Y. Wang, and R. Xiong, "Efficient object manipulation to an arbitrary goal pose: Learning-based anytime prioritized planning," in *ICRA*, 2022, pp. 7277–7283.

[28] B. Calli *et al.*, "Yale-cmu-berkeley dataset for robotic manipulation research," *Int. J. Robot. Res.*, vol. 36, no. 3, pp. 261–268, 2017.

[29] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "Foundationpose: Unified 6d pose estimation and tracking of novel objects," in *CVPR*, 2024, pp. 17 868–17 879.

APPENDIX I

The Direct Prediction (D) method (Fig. 6(a.i)) models the shared grasp distribution directly. It uses a unified EBM E_{ϕ_s} to assign low energy to grasp candidates $(\boldsymbol{g}, w)$ that are feasible under both the initial and goal poses. The method takes as input the combined tuple $(\boldsymbol{t}_{\text{init}}, \boldsymbol{t}_{\text{goal}}, \boldsymbol{g}, w)$ and is trained using the same contrastive objective as the feasibility EBM. During inference, the trained model outputs energy scores for candidate grasps, and a threshold h'_s is applied to classify whether a candidate belongs to the shared grasp set[FT5]:

$$\mathcal{G}_{\text{shared}} = \{(\boldsymbol{g}, w) \mid E_{\phi_s}(\boldsymbol{T}_{\text{init}}, \boldsymbol{T}_{\text{goal}}, \boldsymbol{g}, w) < h'_s\}. \quad (14)$$

The D method is straightforward but requires collecting supervision from explicitly labeled shared grasps and relies on the model's ability to learn the implicit joint constraint.

The Logical Conjunction (L) method (Fig. 6(a.ii)) classifies shared grasps by explicitly enforcing feasibility thresholding at both the initial and goal poses. A candidate is considered shared if it is simultaneously feasible. The method involves two times of thresholding using h_f.

APPENDIX II

Our current dataset representation does not explicitly incorporate object geometries, but only rely on IK feasibility and collision outcomes. Consequently, the transferability observed from Bottle to Mug in Table VII primarily stems from their similar feasible grasp distributions.

TABLE VIII: Model generalization under varying conditions

	Mug			Cracker box			Glass		
	P	R	F1	P	R	F1	P	R	F1
J-1	32.6	72.7	45.0	34.7	45.8	39.4	0.09	84.5	17.0
J-2	17.8	90.5	29.8	13.7	87.7	23.2	0.09	86.8	16.9
J-3	47.8	36.6	41.5	37.1	53.8	43.9	0.09	78.2	16.8

Note During dataset generation, we prepared two versions: one with collision filtering and one without. J-1: No object shape information, no collision information; J-2: With object shape information, with collision information; J-3: No object shape information, with collision information.

To further examine the impact of explicit object geometry, we conducted preliminary studies by incorporating shape embeddings. Specifically, we employed DeepSDF[FT6] to compress object meshes into a latent code z and included it in the EBM input as $E_{\phi_f}(\boldsymbol{T}, \boldsymbol{g}, w, z)$. For training, we used six objects (Bottle, Pitcher, Drill, Milkbox, Bowl, and Bunny). For evaluation, we tested the model on three unseen objects (Mug, Cracker box, and Glass). The results in Table VIII show that adding shape embeddings did not improve performance with the current dataset size. We attribute this to the limited training scale and the complexity of the search space. Meaningful improvements would require intensified training using richer shape variations and is beyond the scope of our current focus.

[FT5] The value of h'_s is selected by maximizing F1 score on a shared grasp validation set constructed as in Section IV.

[FT6] https://github.com/maurock/DeepSDF, pre-trained on 30+ YCB objects