Shared Grasp Estimation to Improve Robotic Pick-and-Place Planning Efficiency

Liang QIN*, Weiwei WAN (The University of Osaka), Jun TAKAHASHI,
Ryo NEGISHI, Masaki MATSUSHITA (H.U. Group Research Institute G.K., Japan),
and Kensuke HARADA (The University of Osaka)

1. Introduction

Grasp selection is crucial for reliable robotic pick-and-place operations. While many approaches focus on object picking, they often neglect downstream placement feasibility, leading to failures, especially in industrial tasks requiring precise alignment and orientation (e.g., inspection, assembly, or retail).

To address this, we propose a learning-based method that predicts grasps with placement awareness, accelerating planning under practical constraints. Our approach builds on the concept of a *shared grasp* [9, 8, 10]—a grasp defined in the object's local frame that remains feasible under both initial and goal poses, shown in Fig. 1. Predicting such grasps reduces computational overhead during the planning process.

We introduce an energy-based model (EBM) that identifies shared grasps under tabletop constraints. Rather than directly learning in joint pose space—which is expensive and data-hungry—we decompose the task into two feasible grasp evaluations and combine their energies. The model assigns low energy to collision-free, IK-feasible grasps and operates over pre-annotated local-frame candidates to improve robustness. Experiments show that our method improves grasp selection efficiency, maintains high success rates, and generalizes to unseen objects and grasps. Our key contributions include:

1. An EBM-based framework for shared grasp prediction using predefined local-frame grasp candidates, ensuring output feasibility and integration with real-world pipelines.
2. A compositional formulation that decomposes shared grasp prediction into two feasible grasp evaluations, improving scalability and generalization in high-dimensional spaces.

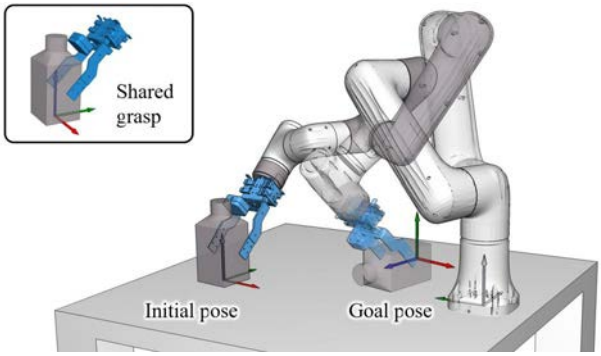


Fig.1 The blue grasp is feasible at both initial and goal poses—a "shared grasp". We propose a method to predict it, thereby accelerating pick-and-place planning.

2. Related Work

2·1 Pick-and-Place Planning

Pick-and-place is a core manipulation task that requires coordinated grasping and placing [1]. This coordination is especially critical in industrial contexts such as assembly [2]. Traditional planning approaches often rely on exhaustive grasp evaluation and motion planning, which become computationally expensive as candidate sets grow. To mitigate this, recent methods leverage learning-based feasibility prediction. Kim et al. [3] used score-space representations to guide planning. Yang et al. [4] introduced sequence-level feasibility prediction using transformers. Park et al. [5] proposed a voxel-based model to evaluate grasp feasibility and accelerate planning.

Our method follows this direction by learning a differentiable energy model over grasp poses conditioned on object configurations. Unlike previous work, we specifically target shared grasp prediction for efficient pick-and-place under geometric constraints.

2·2 Grasp Selection and Regrasp Planning

Grasp selection focuses on identifying feasible grasps that comply with object-specific constraints. Geometry-matching methods [6] and learning-based retrieval [7] are two common strategies. Recent methods have also explored learning grasp feasibility via physical simulation or grasp demonstrations [8]. In shared or regrasp settings, grasps must remain feasible across multiple object poses. Early studies such as [9, 8] explored this idea. Xu et al. [10] predicted grasp distributions for multiple poses, while Cheng et al. [11] learned to place before regrasping.

Our work builds on these ideas and proposes an EBM-based compositional formulation for shared grasp prediction. Inspired by recent energy modeling advances [12], we construct a unified framework that balances efficiency and generalization. The method is validated across large-scale datasets and unseen object categories [13].

3. Modeling Feasible Grasps Using EBM

Traditional pick-and-place planning involves evaluating grasp candidates for feasibility and executing motion planning based on selected grasps. However, as the number of candidates increases, the overall computational burden becomes significant. Inspired by Park et al. [5], we adopt a learning-based approach to predict shared grasps—defined as those feasible under both initial and goal object poses—to accelerate planning. Given object poses $\boldsymbol{T}_{\text{init}}, \boldsymbol{T}_{\text{goal}} \in SE(3)$ and a candidate grasp set $\boldsymbol{\mathcal{G}}_0 = \{(\boldsymbol{g}, w)\}$, where $\boldsymbol{g} \in SE(3)$ and $w \in \mathbb{R}$ denote the gripper pose and width in the object's frame, our goal is to predict the subset $\boldsymbol{\mathcal{G}}_{\text{shared}}$ such that grasp feasibility holds under both poses. A grasp is considered shared if its transformed poses satisfy both IK and collision constraints.

We adopt a two-stage strategy: first, we model single-pose grasp feasibility; second, we combine predictions across poses to determine shared grasps. We use an Energy-Based Model (EBM) to model feasible grasps. The EBM assigns lower energy to feasible $(\boldsymbol{T}, \boldsymbol{g}, w)$ tuples using a learnable function E_{ϕ_f}. The normalized joint distribution is:

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

We implement $E_{\phi_f}(\boldsymbol{T}, \boldsymbol{g}, w)$ as a neural network and train it using negative log-likelihood loss:

$$\mathcal{L}_{\text{nll}} = \mathbb{E}_{\text{pos}}\left[\frac{E_{\phi_f}}{t}\right] + \log Z_{\phi_f}, \tag{2}$$

with contrastive loss:

$$\mathcal{L}_{\text{con}} = \mathbb{E}_{\text{pos}}\left[\frac{E_{\phi_f}}{t}\right] - \mathbb{E}_{\text{neg}}\left[\frac{E_{\phi_f}}{t}\right] \tag{3}$$

and regularization:

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{\text{pos}}\left[\left(\frac{E_{\phi_f}}{t}\right)^2\right] + \mathbb{E}_{\text{neg}}\left[\left(\frac{E_{\phi_f}}{t}\right)^2\right] \tag{4}$$

The total loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{nll}} + \mathcal{L}_{\text{con}} + \alpha\mathcal{L}_{\text{reg}} \tag{5}$$

During inference, grasps are classified as feasible if energy E_{ϕ_f} is below a learned threshold h_f, selected by maximizing the F1 score over the validation set.

For shared grasp prediction, we use the compositional EBM:

$$p(\boldsymbol{T}_{\text{init}}, \boldsymbol{T}_{\text{goal}}, \boldsymbol{g}, w) \propto \exp\left(-\left[E_{\phi_f}(\boldsymbol{T}_{\text{init}}, \boldsymbol{g}, w) + E_{\phi_f}(\boldsymbol{T}_{\text{goal}}, \boldsymbol{g}, w)\right]\right). \tag{6}$$

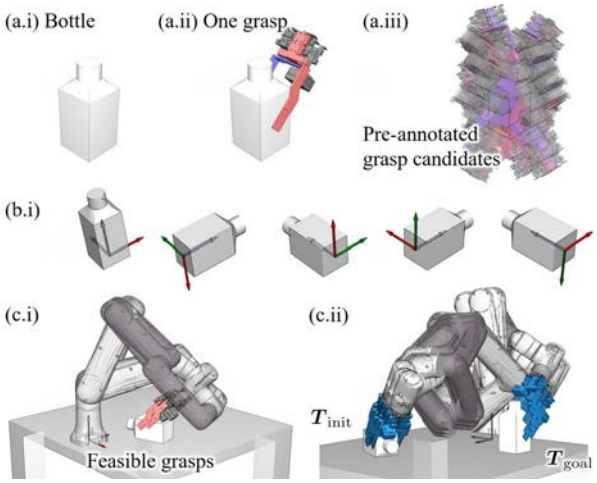


Fig.2 Grasp dataset collection process. (a.i) Target bottle object. (a.ii) The collision-free grasp candidate was generated by sampling the parallel meshes part of the object. (a.iii) The grasp candidate set $\boldsymbol{\mathcal{G}}_0$. (b.i) Stable placements of the object before considering yaw rotations. (c.i) Feasible grasps. (c.ii) Shared grasps.

Shared grasps are predicted via:

$$\boldsymbol{\mathcal{G}}_{\text{shared}} = \left\{(\boldsymbol{g}, w) \mid E_{\phi_f}(\boldsymbol{T}_{\text{init}}, \boldsymbol{g}, w) + E_{\phi_f}(\boldsymbol{T}_{\text{goal}}, \boldsymbol{g}, w) < h_s\right\} \tag{7}$$

We generate training data following [8], using antipodal sampling and collision checking for grasp generation and tabletop-constrained stable placements for object poses. Ground-truth shared grasps are synthesized by intersecting feasible grasp sets under sampled pose pairs.

Fig. 3(a.i) illustrates the shared grasp prediction pipeline using batch energy evaluation and thresholding, and Fig. 3(a.ii)-(a.iii) present the baseline methods. Our approach supports fast and scalable reasoning in pick-and-place tasks.

4. Experiments

We use a 6-DOF Dobot Nova2 arm with a two-finger gripper. Training data is collected in simulation (Fig. 1) with workspace constraints $x \in [-0.45, 0.45]$ m, $y \in [0.1, 0.6]$ m, and $\theta \in [0, 2\pi]$ rad. Resolution is 1 mm for position and 0.01 rad for orientation. The model is trained on an Intel i9-13900KF, 64 GB RAM, and an RTX 4090 GPU. The EBM uses a 3-layer MLP with SELU activations, batch size 1024, learning rate 1×10^{-3}, temperature 0.5, and $\alpha = 0.2$.

We conduct two sets of experiments of the proposed method J and baseline methods D and L(shown in Fig. 3): (1) performance comparison among different prediction methods, and (2) generalization to unseen grasps and objects.

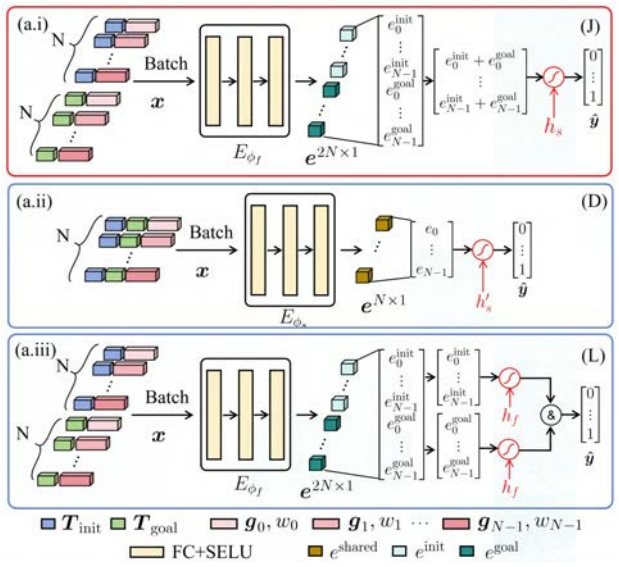


Fig.3 (a.i) Proposed method aggregates EBM energies across poses and classifies shared grasps. (a.ii) Direct Prediction employs a single EBM to estimate shared grasp energy. (a.iii) Logical Conjunction predicts per-pose feasibility with EBM and applies logical AND.

4·1 Comparison with Analytical Methods

We compare the proposed joint estimation method (J method) against analytical baselines using a bottle-shaped object (Fig. 2). Grasp sets of size 57, 109, and 352 are evaluated. Each dataset includes 75k feasible grasps: 50k for training, 15k for testing, and 10k for validation. We test two J variants: J^R (random pick from positives) and J^O (lowest energy pick). Analytical baselines include R (random sampling) and A (IK + collision filtering).

Table 1 Comparison with analytical methods.

	57 candidates			109 candidates			352 candidates		
	A	J^R	J^O	A	J^R	J^O	A	J^R	J^O
$\overline{S_G}$	100	86.6	92.7	100	87.4	88.4	100	81.1	89.9
$\overline{T}$	5.7	1.7		11.0	3.1		34.1	9.2	
$\overline{S_M}$	68.8	63.9	69.9	61.8	63.2	67.1	58.1	59.3	66.0
$\overline{N}$	8.5	12.0	18.2	14.9	15.5	19.1	12.5	12.6	20.1

Note 1 A – Random selection from the intersection of grasps that are IK-feasible and collision-free at both initial and goal poses;
Note 2 J^R – Random selection from the result of J;
Note 3 J^O – Selection of the minimum-energy grasp from the result of J;

Table 1 reports performance. Learning-based methods are significantly faster while maintaining high success rates. With motion planning, J^O yields the highest success rate and lowest average attempts, likely due to energy smoothness aiding planning.

4·2 Generalization Performance

4·2.1 Unseen Grasps

We test whether models generalize to new grasp poses by evaluating on a 922-grasp candidate set. Training is done on grasp sets of size 57–352, using 75k samples per set. Results in Table 2 show that performance improves with more training candidates. The J method consistently performs best. The D method suffers from the insufficiency of training data, and the L method lacks refinement of the threshold under twice-strict classifications.

Table 2 Generality to unseen grasp poses.

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

Note 1 Model 1 ∼ 3 trained on 57, 109, 352 grasp candidates;
Note 2 The best values for shared grasp prediction are highlighted in lime.

4·2.2 Unseen Objects

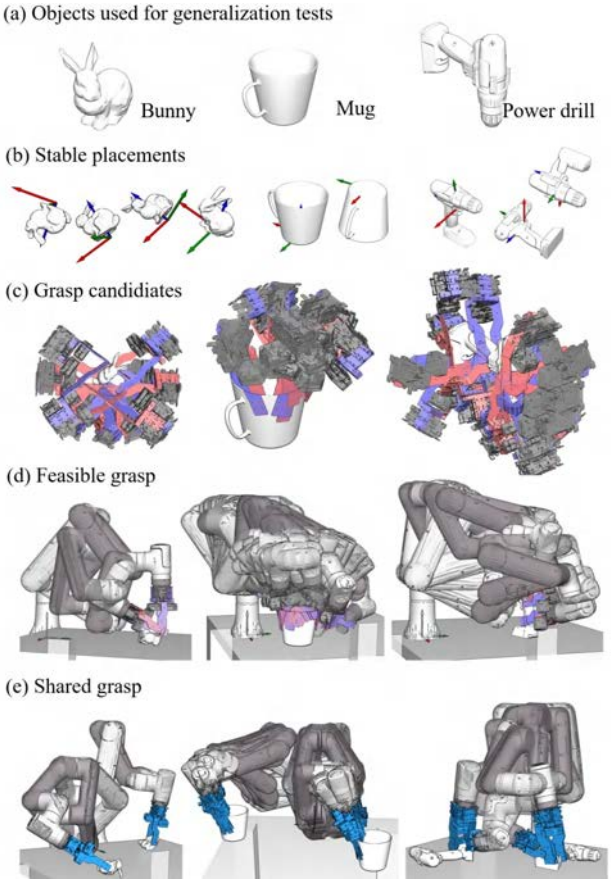


Fig.4 Feasible and shared grasp dataset collection for diverse objects.

We test generalization to unseen object shapes: bottle, bunny, mug, and drill. For each, we collect 150k annotated samples using 180 grasp candidates (Fig. 4).

Table 3 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 (in-distribution) objects; Black: Results of unseen objects.

As shown in Table 3, all methods generalize reasonably well to the mug object. This is likely because its geometry closely resembles the bottle seen during training, leading to similar distributions of feasible grasps. In contrast, generalizing to geometrically dissimilar objects, such as the bunny and drill, proves more difficult.

Among the three methods, D performs the poorest on unseen objects, indicating that directly predicting shared grasps is less robust when training data is limited. While L achieves strong results on in-distribution objects, its generalization to the mug degrades when trained jointly with dissimilar shapes—likely due to cross-domain interference that hampers threshold adaptation. J, although slightly less precise, achieves more consistent generalization across object types, suggesting it is better suited for scaling to a broader range of categories.

5. Future Work and Conclusions

We proposed a shared grasp prediction method based on compositional energy modeling, which efficiently estimates grasps feasible under both initial and goal poses. Compared to analytical methods, our approach significantly reduces planning time while maintaining high accuracy. It also generalizes well to unseen grasps and objects, outperforming direct and logic-based baselines in both efficiency and robustness. Future work will explore extending the method beyond tabletop settings to more complex environments and multi-step manipulation tasks.

References

[1] Shanthi, M. D., & Hermans, T. Pick and Place Planning is Better Than Pick Planning Then Place Planning. *IEEE Robotics and Automation Letters*, **9**(3), 2790–2797 (2024)

[2] Chen, H., Wan, W., Koyama, K., & Harada, K. Planning to Build Block Structures With Unstable Intermediate States Using Two Manipulators. *IEEE Trans. on Automation Science and Engineering*, **19**(4), 3777–3793 (2022)

[3] Kim, B., Wang, Z., Kaelbling, L. P., & Lozano-Pérez, T. Learning to guide task and motion planning using score-space representation. *Int. J. of Robotics Research*, **38**(7), 793–812 (2019)

[4] Yang, Z., Garrett, C., Lozano-Perez, T., Kaelbling, L., & Fox, D. Sequence-Based Plan Feasibility Prediction for Efficient Task and Motion Planning. *Robotics: Science and Systems (RSS)* (2023)

[5] Park, S., Kim, H. C., Baek, J., & Park, J. Scalable learned geometric feasibility for cooperative grasp and motion planning. *IEEE Robotics and Automation Letters*, **7**(4), 11545–11552 (2022)

[6] Herzog, A., Pastor, P., Kalakrishnan, M., Righetti, L., Asfour, T., & Schaal, S. Template-based learning of grasp selection. *IEEE Int. Conf. on Robotics and Automation (ICRA)*, pp. 2379–2384 (2012)

[7] Chen, H., Kiyokawa, T., Wan, W., & Harada, K. Category-association based similarity matching for novel object pick-and-place task. *IEEE Robotics and Automation Letters*, **7**(2), 2961–2968 (2022)

[8] Wan, W., Mason, M. T., Fukui, R., & Kuniyoshi, Y. Improving regrasp algorithms to analyze the utility of work surfaces in a workcell. *IEEE Int. Conf. on Robotics and Automation (ICRA)*, pp. 4326–4333 (2015)

[9] King, J. E., Klingensmith, M., Dellin, C. M., Dogar, M. R., Velagapudi, P., Pollard, N. S., & Srinivasa, S. S. Pregrasp Manipulation as Trajectory Optimization. *Robotics: Science and Systems (RSS)* (2013)

[10] Xu, P., Chen, Z., Wang, J., & Meng, M. Q.-H. Learning to predict diverse stable placements for extrinsic manipulation on a support plane. *IEEE Trans. on Cognitive and Developmental Systems*, **16**(3), 1095–1107 (2023)

[11] Cheng, S., Mo, K., & Shao, L. Learning to Regrasp by Learning to Place. *Conf. on Robot Learning (CoRL)*, pp. 277–286 (2022)

[12] Du, Y., Li, S. & Mordatch, I. Compositional visual generation with energy-based models. *Advances in Neural Information Processing Systems*, **33**, 6637–6647 (2020)

[13] Xu, K., Zhou, Z., Wu, J., Lu, H., Xiong, R., & Wang, Y. Grasp, See, and Place: Efficient Unknown Object Rearrangement With Policy Structure Prior. *IEEE Transactions on Robotics*, **41**, 464–483 (2025)