ReorientDiff: Diffusion Model based Reorientation for Object Manipulation

Utkarsh A. Mishra and Yongxin Chen

Abstract— **The ability to manipulate objects in a desired configurations is a fundamental requirement for robots to complete various practical applications. While certain goals can be achieved by picking and placing the objects of interest directly, object reorientation is needed for precise placement in most of the tasks. In such scenarios, the object must be reoriented and re-positioned into intermediate poses that facilitate accurate placement at the target pose. To this end, we propose a reorientation planning method, ReorientDiff, that utilizes a diffusion model-based approach. The proposed method employs both visual inputs from the scene, and goal-specific language prompts to plan intermediate reorientation poses. Specifically, the scene and language-task information are mapped into a joint scene-task representation feature space, which is subsequently leveraged to condition the diffusion model. The diffusion model samples intermediate poses based on the representation using classifier-free guidance and then uses gradients of learned feasibility-score models for implicit iterative pose-refinement. The proposed method is evaluated using a set of YCB-objects and a suction gripper, demonstrating a success rate of 96.5% in simulation. Overall, our study presents a promising approach to address the reorientation challenge in manipulation by learning a conditional distribution, which is an effective way to move towards more generalizable object manipulation. For more results, checkout our website: https://utkarshmishra04.github.io/ReorientDiff.**

I. Introduction

Rearranging objects in a desired pose is an important skill necessary for daily activities at home as well as for specific arrangement and packing applications in the industry. Performing such a task requires extracting object information from visual-sensor data and planning a pick-place sequence [1], [2]. While a single-step pick-place sequence is a viable solution, placing the object at a specific position and orientation is not always feasible. Reorientation is a helpful strategy when successfully changing an object's pose allows its placement at the target pose [3]. Reorientation ensures feasible intermediate transition poses in scenarios where there are no common grasps between the current pose and an object's desired placement pose.

In classical approaches, such a problem is usually tackled by using trajectory planners [4] to plan motion from the current pose to the desired pose via diverse candidate intermediate poses. Such an exhaustive search is expensive on time and is limited by choice of the number of intermediate pose options. Recently, Wada *et al.* [3] proposed a data-driven sampling-based solution to reorientation using learned models that predict the feasibility score of an intermediate pose. While their method significantly improved the success rate and planning time, the approach relied on the target object's specification and placement pose. Lately, with the advances in language descriptor foundation models like CLIP [5], which projects images and texts to a common feature space, such specifications can be directly correlated between visual information and suitable language commands, thus empowering human-robot interaction. This motivated us to explore grounding the problem statement of reorientation on language and hence embed semantic knowledge of the task with the spatial structure of the scene [6].

In this paper, we propose ReorientDiff, a diffusion model-based reorientation pose generation pipeline for solving the task proposed by [3] for picking objects from a cluttered pile and placing it in the target pose specified through language descriptions. The core idea of our approach is to visualize the feasible intermediate poses as distributions. Such a distribution can be captured by a diffusion model and will be conditioned on the object's current and target pose, or in a more general multi-object scenario, on the pile of pickable objects and the occupancy of the target location. Note that diffusion models have also been successfully used for motion planning [7], [8], grasp planning [9], and object rearrangement [10] applications.

To enable interaction using natural language directly, we use pre-trained CLIP embeddings with an object segmentation model to structure object selection, pose prediction, and target object segmentation networks for the task. Considering the intermediate features as a generic scene and target representation in reduced dimensionality, the diffusion model samples reorientation poses conditioned on such features, which are further implicitly refined by a feasibility-score-based discriminator similar to the models used by [11], [3]. We combine a generic classifier-free conditional sampling [12] with classifier-guided sampling [13] to sample from diffusion models. For the tasks, we consider reorientation of objects in the YCB dataset [14] that are feasible for suction grippers. For each selected object, we choose target locations on multiple shelf levels and four possible target orientations. Our method samples reorientation poses in continuous space without any discretization or candidate pose selection and reaches a cumulative success rate of 96.5% as evaluated on selected individual objects.

II. Related Work

A. Object Manipulation: Pick and Place

While traditional methods have tried to solve the pick-and-place task using grasp planning [15], [16], [17] with known

Utkarsh A. Mishra and Yongxin Chen are affiliated to the Institute for Robotics and Intelligent Machines (IRIM), Georgia Institute of Technology umishra31@gatech.edu, yongchen@gatech.edu

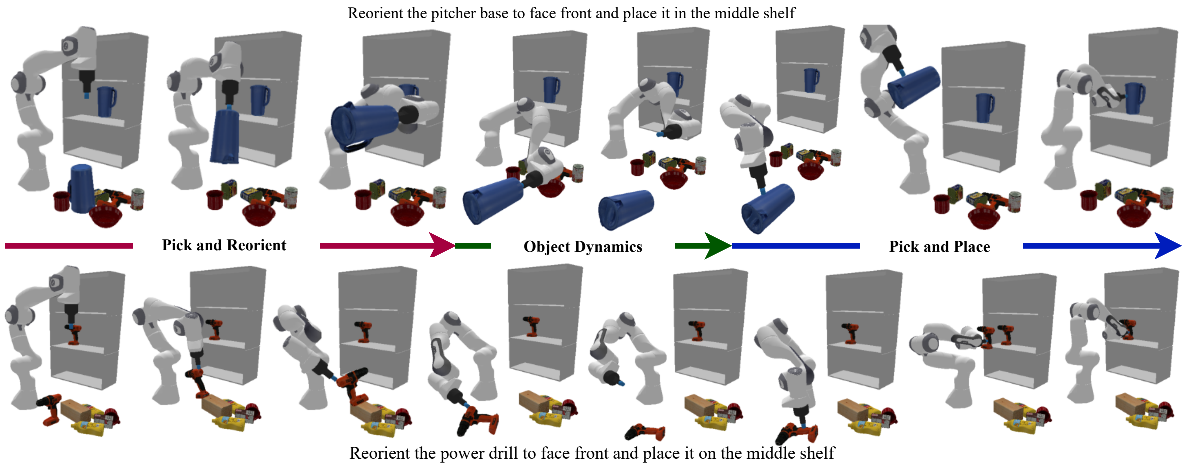

Reorient the pitcher base to face front and place it in the middle shelf
Pick and Reorient
Object Dynamics
Pick and Place
Reorient the power drill to face front and place it on the middle shelf

Fig. 1: **Reorientation for precise target placement** The above figure represents the phenomenon of reorientation in which an object from a cluttered file has to be placed precisely in a shelf (target position shown). As the object cannot be directly place at the target location, our proposed method, ReorientDiff, samples a reorientation pose using a learned conditional distribution by a diffusion model. Such a proposed reorientation pose acts as a transition for facilitating successful placement. We also consider and take advantage of the object dynamics, as introduced by Wada *et al.* [3], by which we ensure that un-grasping an object in an unstable pose will eventually allow the object to settle at some favourable pose.

object geometries or using pose estimation methods [18], [19], [20], the recent literature has focused more on vision-based object manipulation [1], [21]. Solving single-step pick and place tasks is typically achieved by planning grasp poses using segmentation and depth maps of the object, where it is considered that a picked object can be placed within the region of interest (like in a box) [22], [23]. Recent studies have also shown object rearrangement planning capabilities [6], [24] where a target location is sampled based on some user-specified goal. Then the whole-pipeline for generating a collision free trajectory from current to target location is planned. Some works have proposed object rearrangement as a long horizon problem [2] consisting of multiple sequential pick and place actions to achieve a desired configuration.

B. Language Models for Robotics

Language models like GPT-2 [25] and GPT-3 [26] have proven to be quite effective in grounding the task's semantics with the scene's spatial features using several foundation models. One such foundation model is CLIP [5] which encodes visual and language information into common representation space and has been helpful in learning policies for generalized pick-place tasks in planar tabletop [6] and 3D [24] manipulation and for control of embodied AI agents [27], [28]. Further, language models have also been used in language-conditioned object rearrangement planning [10], [29] and supplying high-level instructions for long-horizon planning [30].

C. Reorientation and Regrasping

Reorientation is a vital capability required for solving complex manipulation tasks. Prior research have explored this direction by planning to reorient objects using extrinsic supports [31], [32], which enables them to re-grasp the object in a desired way. While [31] proposed a graph neural network structure for pose sequencing and [32] used an end-to-end point-cloud based model for predicting reorientation poses, [33] proposed a heuristic based method for reorienting rock structures in excavation. Recently, ReorientBot [3] was proposed to solve the reorientation task using learned feasibility prediction models and rejection sampling.

D. Generative Models for Robotics

Generative models like VAE have been used for planning grasps [11] using visible point-cloud of objects and for constructing embedding space for high-level tasks for various downstream planning. Recently, diffusion models have been used extensively in literature for trajectory planning from imitation data [7], [8] and for generating target poses for language-conditioned object rearrangement tasks [10]. With language-guided scene and video generation applications, such models have been used for generating task-videos for robot learning [34] and generalizing to unseen scenarios [35].

III. Preliminary: Diffusion Models

Consider samples x_0 from an unknown data distribution $q(x_0)$; diffusion models [36] learn to estimate the distribution by a parameterized model $p_\theta(x_0)$ using the given samples. The procedure is completed in two steps: the forward and the reverse diffusion processes. The former continuously injects Gaussian noise in x_0 to create a Markov chain with latents $x_{1:K}$ following transitions:

$$q(x_{1:K}|x_0) = \prod_{k=1}^{K} q(x_k|x_{k-1}), \tag{1}$$

where $q(x_k|x_{k-1}) = \mathcal{N}(x_k; \sqrt{1-\beta_k}x_{k-1}, \beta_k\mathbf{I})$ is the per-step noise injection following variance schedule $\beta_1, \dots, \beta_K$. This leads to the distribution $q(x_k|x_0) = \mathcal{N}(x_k; \sqrt{\bar{\alpha}_k}x_0,\ (1-\bar{\alpha}_k)\ \mathbf{I})$ following notations introduced in [37] as $\alpha_k = 1 - \beta_k$ and $\bar{\alpha}_k = \prod_{i=1}^{k} \alpha_i$. Note that $\bar{\alpha}_K \approx 0$ and thus $x_K \sim \mathcal{N}(0, \mathbf{I})$. The reverse diffusion learns to denoise the data starting from

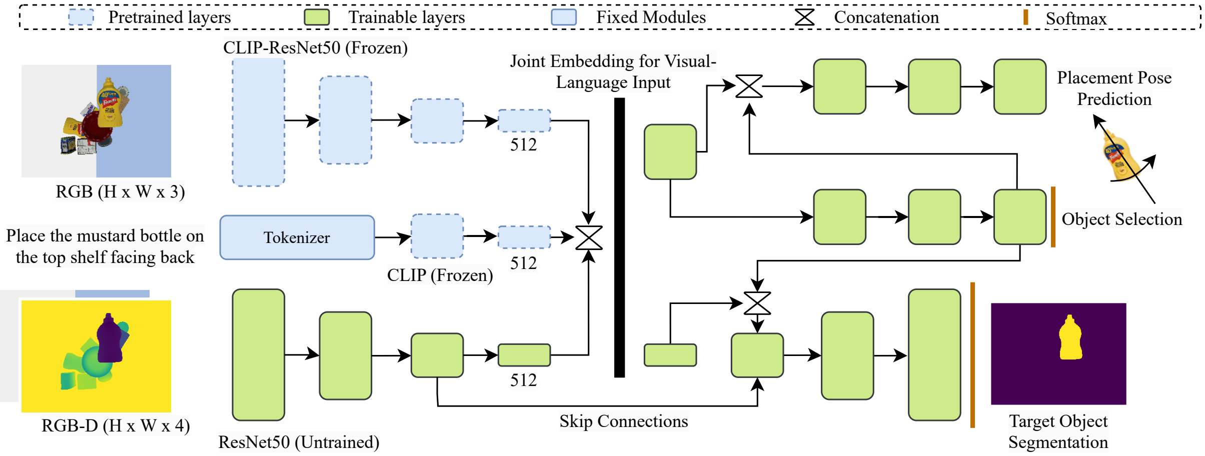


Fig. 2: **Joint Embedding Construction** We use a pre-trained CLIP-ResNet50 and BPE-based Tokenizer with CLIP language model for obtaining a semantic embedding of the tabletop RGB image and instruction prompt, respectively. While keeping CLIP layers frozen, we train another ResNet50 encoder for spatial RGB-D observations and combine them with the semantic embeddings to obtain joint embeddings for visual-language inputs. We train these latent representations with respect to the object of interest, placement pose, and object location (segmentation) predictions. It is worth noting that the predicted object information is also used for predicting the placement pose and the target object segmentation. Further, the addition of skip-connection also ensures that the segmentation map construction is accurate while filling up the embedding vector with only the necessary information. The proposed pipeline shown above creates a latent space that is consistent with the three aspects of interest by minimizing information loss.

x_K and following $p_\theta(x_{k-1}|x_k) = \mathcal{N}(x_{k-1}; \mu_\theta(x_k, k), \beta_k \mathbf{I})$ where

$$\mu_\theta(x_k, k) = \frac{1}{\sqrt{\alpha_k}}\Big(x_k - \frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}}\epsilon_\theta(x_k, k)\Big). \tag{2}$$

The parameterized model $\epsilon_\theta(x_k, k)$ is called the score-function, and it is trained to predict the perturbations and the noising schedule by the score-matching objective [38]

$$\arg\min_\theta \mathbb{E}_{x_0 \sim q, \epsilon \sim \mathcal{N}(0,\mathbf{I})}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_k}x_0 + \sqrt{1-\bar{\alpha}_k}\epsilon)\|^2\right] \tag{3}$$

In particular, such a score function represents the gradients of the learned probability distribution as

$$\nabla_{x_k} \log p_\theta(x_k) = -\frac{1}{\sqrt{1-\bar{\alpha}_k}}\epsilon_\theta(x_k, k). \tag{4}$$

IV. Reorientation

Following the previous environment setup by Wada *et al.* [3], we construct the reorientation scenario as a task of i) selecting an object of interest from a pile of cluttered objects, ii) calculating feasible grasp poses for picking, iii) calculating grasp poses for placement with prior knowledge of the mesh of the selected object and iv) finding suitable reorientation poses using our proposed pipeline based on diffusion models. This section describes the pipeline for creating a generic scene and task embedding space, followed by generating grasp poses and training the feasibility score models.

A. Constructing Generic Scene-Task Representations

We define a scene as the location and occupancy of the place from where a target object should be picked and a task as the language prompt containing the descriptions for selecting the target object and deciding placement poses. A top-down RGB-D camera provides an image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$ and a heightmap $\mathcal{H} \in \mathbb{R}^{H\times W\times 1}$ as the description of the pile. Motivated from previous work [6], [24] on learning semantic and spatial embeddings, we use pre-trained CLIP foundation model for obtaining semantic embeddings from the image $\mathcal{I}$ and language $\mathcal{L}$, and combine them with spatial embeddings for target object segmentation to get a joint embedding Φ as generic scene-task representation as shown in Fig. 2. The embedding is further used to predict the target object as a one-hot vector of all the objects of interest and the final placement pose.

B. Sampling Grasp Poses

We generate grasp poses by following the classical approach of converting the heightmap into a point cloud representation and eventually to a point-normal representation. The predicted target object segmenatation of the scene is then used to obtain the surface normals of the target object. After performing an edge masking using the Laplacian of the surface normals, the remaining point-normals on the surface are feasible grasp poses. While we sample grasp poses η_1 for picking the object from the pile in the aforementioned manner, we assume that we have the mesh of the selected object for sampling grasp poses η_2 for placing the object at the predicted pose.

C. Training Feasibility Score Models

Following prior works [11], [3], [29], a feasibility prediction model is important for early-evaluation and rejection of unfavourable samples. Such a feasibility model predicts the probability of success of a given grasp pose in successfully grasping an object in some candidate pose for a specified scene representation. The phenomenon of grasp success evaluation in dynamic reorientation pose, as addressed by [3], is of particular interest for our setup. Modelling dynamics for every object is indeed non-trivial and adds to the complexity; hence the feasibility model implicitly takes care of the dynamics of the object after deactivating the grasp. For checking feasibility or the probability of success (y) of

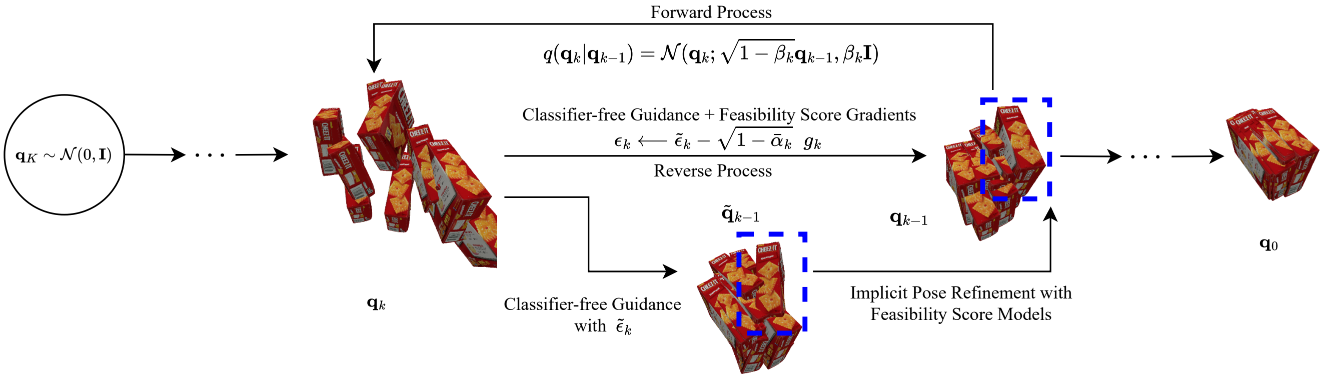

Forward Process
$q(\mathbf{q}_k|\mathbf{q}_{k-1}) = \mathcal{N}(\mathbf{q}_k; \sqrt{1-\beta_k}\mathbf{q}_{k-1}, \beta_k\mathbf{I})$
$\mathbf{q}_K \sim \mathcal{N}(0, \mathbf{I})$
Classifier-free Guidance + Feasibility Score Gradients
$\epsilon_k \longleftarrow \tilde{\epsilon}_k - \sqrt{1-\bar{\alpha}_k}\ g_k$
Reverse Process
$\mathbf{q}_k$
$\tilde{\mathbf{q}}_{k-1}$
$\mathbf{q}_{k-1}$
$\mathbf{q}_0$
Classifier-free Guidance with $\tilde{\epsilon}_k$
Implicit Pose Refinement with Feasibility Score Models

Fig. 3: **Forward and Reverse Diffusion Process** The above figure shows the forward diffusion and the reverse denoising and sampling process of ReorientDiff. As described in Section V, following classifier-free guidance will result in high-likelihood samples with high-variance in terms of success feasibility of the samples. Using the feasibility score gradients, we realize an implicit iterative pose refinement, as marked by the blue box in the figure. This significantly decrease variance and ensure high success feasibility of the samples.

sampled grasps for candidate reorientation poses $\mathbf{q}$, we train two models:

- For predicting success of reorientation from the current pose in a pile to a candidate pose given pick grasp poses (η_1) and scene representation, denoted as $\mathcal{M}_1(y|\eta_1, \mathbf{q}, \Phi)$
- For predicting success of post-grasp deactivation pose from the candidate pose and placement grasp poses (η_2), denoted as $\mathcal{M}_2(y|\eta_2, \mathbf{q}, \Phi)$

V. ReorientDiff: Diffusion for Reorientation

We aim to generate intermediate reorientation poses for the target object, which enables successive placement at the desired pose and is reachable from the current pose. We introduce a diffusion model based approach to sample most probable successful reorientation poses ($\mathbf{q}$) conditioned on the scene representation priors (Φ), denoted as $p(\mathbf{q}|\Phi)$, which already contains the spatial and semantic information about the scene and the task. The denoising process can be further flexibly conditioned by sampling from modified distributions of the form

$$p_h(\mathbf{q}) \propto p(\mathbf{q}|\Phi)h(\mathbf{q}, \Phi), \tag{5}$$

where $h(\mathbf{q}, \Phi)$ can represent several grasp success probability heuristics. By separating the grasp success from reorientation candidate sampling, the diffusion model trained for reorientation poses can be reused for varied selection of picking (η_1) and placement grasp poses (η_2).

A. Classifier-free Conditional Pose Generation

Following the distribution defined in (5), we use classifier-free guidance [12] to sample high-likelihood reorientation poses for a particular scene-task representation. We train a score-network [38], $\epsilon_\theta(\mathbf{q}_k, \Phi) \propto \nabla_{\mathbf{q}_k} \log p(\mathbf{q}_k|\Phi)$, to denoise from $\mathbf{q}_K \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to possible reorientation poses $\mathbf{q}_0$ from a K-step reverse diffusion denoising process. For each step, we calculate $\tilde{\epsilon}_k$ as

$$\tilde{\epsilon}_k = \epsilon_\theta(\mathbf{q}_k, \Phi) + w_c\Big(\epsilon_\theta(\mathbf{q}_k, \Phi) - \epsilon_\theta(\mathbf{q}_k, \emptyset)\Big) \tag{6}$$

The scalar w_c implicitly guides the reverse-diffusion towards poses that best satisfy the scene-task representations. Further, we calculate the successive samples for the next $(k-1)^{th}$ step using the DDIM [37] sampling strategy and $\tilde{\epsilon}_k$ as follows:

$$\tilde{\mathbf{q}}_{k-1} \longleftarrow \sqrt{\bar{\alpha}_{k-1}}\Big(\frac{\mathbf{q}_k - \sqrt{1-\bar{\alpha}_k}\ \tilde{\epsilon}_k}{\sqrt{\bar{\alpha}_k}}\Big) + \sqrt{1-\bar{\alpha}_{k-1}}\ \tilde{\epsilon}_k \tag{7}$$

where, $\bar{\alpha}_k$ is as described in Section III.

B. Feasibility Guided Pose Refinement

We use the two feasibility-score prediction models ($\mathcal{M}_1$ and $\mathcal{M}_2$), which are pre-trained for predicting grasp feasibility for picking grasp, reorientation pose pairs and placement grasp, reorientation pose pairs, respectively. In such a case, the scores can be converted into probability distributions for each heuristic, defined as, for each $i = 1, 2$,

$$h_i \equiv p(y=1|\eta_i, \mathbf{q}, \Phi)|_{\mathcal{M}_i} = \exp\Big(-(1 - \mathcal{M}_i(y|\eta_i, \mathbf{q}, \Phi))^2\Big)$$

Following classifier-based guidance [13] formulation for the heuristics, the reverse diffusion can be formulated as:

$$\begin{aligned} &p_h(\mathbf{q}_k|\mathbf{q}_{k+1}, y, \Phi) \propto \\ &p(\mathbf{q}_k|\mathbf{q}_{k+1}, \Phi)\ \ p(y|\eta_1, \hat{\mathbf{q}}_0^k, \Phi)|_{\mathcal{M}_1}\ p(y|\eta_2, \hat{\mathbf{q}}_0^k, \Phi)|_{\mathcal{M}_2} \end{aligned} \tag{8}$$

where, $\hat{\mathbf{q}}_0^k$ is the sample proposed at diffusion step k and defined as:

$$\hat{\mathbf{q}}_0^k = \frac{\mathbf{q}_k - \sqrt{1-\bar{\alpha}_k}\ \tilde{\epsilon}_k}{\sqrt{\bar{\alpha}_k}} \tag{9}$$

Considering Taylor first order approximations for heuristics and standard reverse process Gaussian $(\mu_\theta(\mathbf{q}_k, k, \Phi), \beta_k\mathbf{I})$ as described in Section III, we get the new mean $(\mu_{\theta,h}(\mathbf{q}_k, k, \Phi))$ for the distribution $p_h(\mathbf{q}_k|\mathbf{q}_{k+1}, y, \Phi)$ in (8) as:

$$\begin{aligned} &\mu_{\theta,h}(\mathbf{q}_k, k, \Phi) \\ &= \mu_\theta(\mathbf{q}_k, k, \Phi) + \beta_k \sum_{i=1}^{2} w_i \nabla_{\mathbf{q}_k} \log p(y|\eta_i, \mathbf{q}_k, \Phi)|_{\mathcal{M}_i} \\ &= \mu_\theta(\mathbf{q}_k, k, \Phi) - \beta_k \sum_{i=1}^{2} w_i \nabla_{\mathbf{q}_k} \Big[1 - \mathcal{M}_i(y|\eta_i, \hat{\mathbf{q}}_0^k, \Phi)\Big]^2. \end{aligned}$$

In view of (2), we then obtain the modified score

$$\epsilon_k \longleftarrow \tilde{\epsilon}_k - \sqrt{1-\bar{\alpha}_k}\ g_k$$

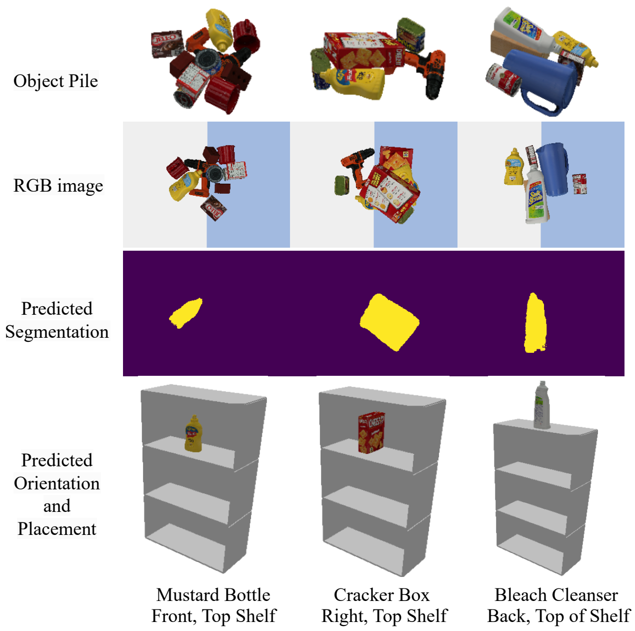


Fig. 4: **Visual Analysis of Scene-Task Network Performance** The scene-task network maps the visual (row 2) image of the pile (row 1) and language (bottom row) inputs to a feature space which is used to predict the placement location (row 4) and target object segmentation (row 3).

where $g_k = -\beta_k \sum_{i=1}^{2} w_i \nabla_{\mathbf{q}_k} \left[1 - \mathcal{M}_i(y|\eta_i, \hat{\mathbf{q}}_0^k, \Phi)\right]^2$. We notice that injecting noise to g_k, as in stochastic DDIM, can slightly improve the performance. We calculate the final $\mathbf{q}_{k-1}$ using the refined ϵ_k in (7). A visual clarification of the forward and reverse diffusion is shown in Fig. 3.

VI. Results: Simulation

Based on the environment setup as discussed in Section IV, we create datasets, train diffusion and feasibility score models and evaluate them in simulation for proper placement conditions.

A. Dataset Generation and Training

We use PyBullet [39] and an OMPL [40] based motion planner to solve for collision-free path between current pose and a candidate reorientation pose and from the reorientation pose to the ground-truth placement pose for diverse set of YCB-objects and target locations. We sampled approximately 40000 candidate poses following Wada *et al.* [3]. The goal properties were converted into modular language instructions, and the success of pick and place for both the steps was recorded. The scene and task properties were used to construct the joint visual-language embedding space, which was further used to train the feasibility score models using binary success labels. Eventually, we train a conditional diffusion model using only the successful reorientation poses. Such a diffusion model is reusable for diverse set of grasp poses based on the feasibility score models.

B. Performance Evaluation: Scene-Task Representation

To evaluate the quality of the scene-task embedding network, we analyze the accuracy of the object selection and placement pose prediction along with the error in the predicted segmentation. We show a visual analysis in Fig. 4 where the output segmentation and the predicted placement pose in the shelf are shown for three scenes and tasks. For accurate shelf-level estimation, we round each object's predicted height to the nearest shelf-level height, and a similar post-processing is conducted for the object orientation. To add complexity, although we consider only four orientations: front, back, left and right, we discretize the possible orientations into 8 possible options and round the predicted orientation to the nearest option.

Numerically, the object selection network was 100% accurate, and the number of pixels wrongly classified was about 1% of the complete image on average over 100 random samples. The average error in predicting the height of the target placement after post-processing is around 8 mm, and the mean error in the yaw angle of the predicted pose is 0.3 rad.

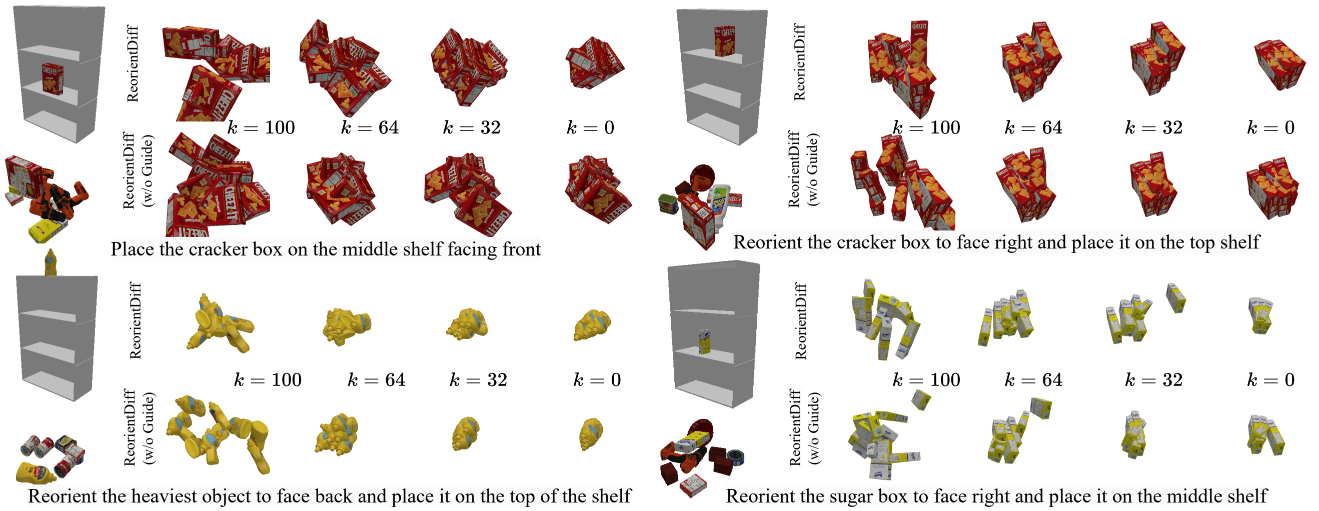
C. Performance Evaluation: Diffusion with Guidance

The trained classifier-free conditional diffusion model and the score feasibility models are used to perform the reverse diffusion using the classifier-free guidance with and without feasibility score guidance. Experiments comparing the performance of both the methods are shown in Fig. 5 for a set of YCB Objects [14] and different scene-task scenarios where only 50 candidate poses are sampled and top 10 high-likelihood poses are selected. The comparison shows that while the classifier-free guidance is good enough to sample high-likelihood reorientation poses, the primary purpose of the feasibility score gradients is to reduce the variance in the pose generation and ensure high success probability. A numerical analysis of the overall success is shown and compared with the rejection sampling based baseline [3] in Table I.

TABLE I: Success evaluation of the proposed method as compared to the rejection sampling based baseline. The ReorientDiff algorithm was tested for more than 300 different scene task settings consisting of equal distribution of the selected objects and all the orientations. A task is considered a success if it is completed at-least once in 4 random seeds.

Method	**Success (%) Reorient**	**Success (%) Place**	**Success (%) Overall**
ReorientBot	97.9	95.1	93.2
ReorientDiff (w/o Guide)	97.4	86.3	85.8
ReorientDiff	**98.9**	**96.5**	**94.8**

The reorientation success percentage holds different relevance as compared to the baseline. The baseline does two step reverse rejection sampling where reorientation search is conducted over candidates which are feasible for placement, so there might be a scenario where there is no solution possible. For the case of ReorientDiff, the reorientation success measures the capability of the diffusion model to generalize to poses which ensure reorientability and scope for future placement. Higher reorientation success and lower placement success is an indication that the model is short-sighted and is giving importance to a single step success metric. From Table I, we ensure high reorientability success along with better placement success, even without any candidate pose discretization. The overall success is based



Place the cracker box on the middle shelf facing front

Reorient the cracker box to face right and place it on the top shelf

Reorient the heaviest object to face back and place it on the top of the shelf

Reorient the sugar box to face right and place it on the middle shelf

Fig. 5: **Reverse Diffusion for Reorientation Pose Generation** The reverse sampling process for 4 k-steps at $k = 100, 64, 32, 0$ for $K = 256$ in four different scene-task scenarios comprising of the Cracker Box, Mustard Bottle and Sugar Box in different target orientations are shown above. The scenes are shown in the left-side of every sub-figure and consists of the pile with the target object and the predicted placement location on the shelf. The language prompt defining each of the tasks is mentioned below each sub-figure. It consists of either the absolute (the object's name) or the relative (heaviest/lightest) reference to the object and details about the target placement.

on the accurate placement of the object from the reoriented pose, and it represents the successful completion of a task. The metric is measured by calculating the difference between the desired and the pose after final placement.

TABLE II: Success evaluation with different levels of discretization while sampling using ReorientDiff.

ReorientDiff K	Success (%) Reorient	Success (%) Place	Success (%) Overall
$K = 256$	**98.9**	**96.5**	**94.8**
$K = 100$	99.3	92.4	91.5
$K = 50$	97.5	91.1	88.6

D. Performance Evaluation: K-Step Reverse Diffusion

Sampling from a trained diffusion models is flexible and can be achieved using different levels of discretization between $x_K \sim \mathcal{N}(0, \mathbf{I})$ to meaningful reorientation poses. We perform the complete analysis for multiple values of the number of reverse denoising steps K as shown in Table II. It is evident that minimizing resolution degrades overall performance, but even with much fewer resolutions, ReorientDiff can reach a descent performance.

Following our analysis on performance, we explored the time taken for overall planning of a successful reorientation pose from a given scene and corresponding task information. We provide the recorded timings for all of our ablations as well as the baseline in Table III.

TABLE III: Computational analysis of the planning time for finding a suitable reorientation pose for the proposed method, ReorientDiff, along with the baseline and all conducted ablations.

Method	Planning Time (sec)
ReorientBot	2.5
ReorientDiff (w/o Guide)	2.7
ReorientDiff @ $K = 256$	**5.3**
ReorientDiff @ $K = 100$	2.5
ReorientDiff @ $K = 50$	1.5

Our findings show that ReorientDiff is computationally heavy due to gradient calculations for reverse denoising steps. Without using the guidance from the feasibility-score models, classifier-free guidance requires similar time as the baseline, ReorientBot. However, as we decrease the discretization resolution, the planning time decreases significantly with some trade-off in performance, as shown in Table II. We believe that using higher-order solvers such as one proposed in [41], similar level of performance as ReorientDiff (w/ $K = 256$) can be achieved at the computation cost of $K = 50$. However, such an analysis is out of scope of the proposed methodology. Hence, from all of our visual and empirical analysis, ReorientDiff successfully proves that formulating the problem of reorientation as learning a conditional distribution is an effective way to move towards more generalizable object manipulation.

VII. Conclusion

Diffusion models are powerful generative models capable of modeling (conditional) distributions. In the proposed method, ReorientDiff exploits the capabilities of such models to predict reorientation poses conditioned on a compact scene-task representation embedding containing information about the target object and its placement location. Further, the samples are refined using learned feasibility-score models to reduce uncertainty and ensure success of the planned intermediate poses. Considering as little as 10 reorientation poses, we achieved an overall success rate of 96.5% across variety of objects. We consider incorporating more efficient sampling schemes and better generalization performance for unseen objects and placement goals as a potential future work.

References

[1] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani, *et al.*, "Transporter networks: Rearranging the visual world for robotic manipulation," in *Conference on Robot Learning*, pp. 726–747, PMLR, 2021.

[2] B. Tang and G. S. Sukhatme, "Selective object rearrangement in clutter," in *6th Annual Conference on Robot Learning*, 2022.

[3] K. Wada, S. James, and A. J. Davison, "Reorientbot: Learning object reorientation for specific-posed placement," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 8252–8258, IEEE, 2022.

[4] J. J. Kuffner and S. M. LaValle, "Rrt-connect: An efficient approach to single-query path planning," in *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065)*, vol. 2, pp. 995–1001, IEEE, 2000.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[6] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Conference on Robot Learning*, pp. 894–906, PMLR, 2022.

[7] M. Janner, Y. Du, J. Tenenbaum, and S. Levine, "Planning with diffusion for flexible behavior synthesis," in *International Conference on Machine Learning*, 2022.

[8] A. Ajay, Y. Du, A. Gupta, J. Tenenbaum, T. Jaakkola, and P. Agrawal, "Is conditional generative modeling all you need for decision-making?," *arXiv preprint arXiv:2211.15657*, 2022.

[9] J. Urain, N. Funk, G. Chalvatzaki, and J. Peters, "Se (3)-diffusionfields: Learning cost functions for joint grasp and motion optimization through diffusion," *arXiv preprint arXiv:2209.03855*, 2022.

[10] W. Liu, T. Hermans, S. Chernova, and C. Paxton, "Structdiffusion: Object-centric diffusion for semantic rearrangement of novel objects," *arXiv preprint arXiv:2211.04604*, 2022.

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

[12] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[13] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.

[14] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks for manipulation research," in *2015 international conference on advanced robotics (ICAR)*, pp. 510–517, IEEE, 2015.

[15] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," *arXiv preprint arXiv:1703.09312*, 2017.

[16] J. Mahler, M. Matl, X. Liu, A. Li, D. Gealy, and K. Goldberg, "Dex-net 3.0: Computing robust vacuum suction grasp targets in point clouds using a new analytic model and deep learning," in *2018 IEEE International Conference on robotics and automation (ICRA)*, pp. 5620–5627, IEEE, 2018.

[17] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg, "Learning ambidextrous robot grasping policies," *Science Robotics*, vol. 4, no. 26, p. eaau4984, 2019.

[18] M. Zhu, K. G. Derpanis, Y. Yang, S. Brahmbhatt, M. Zhang, C. Phillips, M. Lecce, and K. Daniilidis, "Single image 3d object detection and pose estimation for grasping," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3936–3943, IEEE, 2014.

[19] A. Zeng, K.-T. Yu, S. Song, D. Suo, E. Walker, A. Rodriguez, and J. Xiao, "Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge," in *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 1386–1383, IEEE, 2017.

[20] X. Deng, Y. Xiang, A. Mousavian, C. Eppner, T. Bretl, and D. Fox, "Self-supervised 6d object pose estimation for robot manipulation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3665–3671, IEEE, 2020.

[21] D. Seita, P. Florence, J. Tompson, E. Coumans, V. Sindhwani, K. Goldberg, and A. Zeng, "Learning to rearrange deformable cables, fabrics, and bags with goal-conditioned transporter networks," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4568–4575, IEEE, 2021.

[22] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," in *2016 IEEE international conference on robotics and automation (ICRA)*, pp. 3406–3413, IEEE, 2016.

[23] A. Zeng, S. Song, K.-T. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. 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*, vol. 41, no. 7, pp. 690–705, 2022.

[24] M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-actor: A multi-task transformer for robotic manipulation," *arXiv preprint arXiv:2209.05451*, 2022.

[25] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[26] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[27] A. Khandelwal, L. Weihs, R. Mottaghi, and A. Kembhavi, "Simple but effective: Clip embeddings for embodied ai," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14829–14838, 2022.

[28] D. Batra, A. Gokaslan, A. Kembhavi, O. Maksymets, R. Mottaghi, M. Savva, A. Toshev, and E. Wijmans, "Objectnav revisited: On evaluation of embodied agents navigating to objects," *arXiv preprint arXiv:2006.13171*, 2020.

[29] W. Liu, C. Paxton, T. Hermans, and D. Fox, "Structformer: Learning spatial structure for language-guided semantic rearrangement of novel objects," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 6322–6329, IEEE, 2022.

[30] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[31] S. Cheng, K. Mo, and L. Shao, "Learning to regrasp by learning to place," in *5th Annual Conference on Robot Learning*, 2021.

[32] P. Xu, Z. Chen, J. Wang, and M. Q.-H. Meng, "Planar manipulation via learning regrasping," *arXiv preprint arXiv:2210.05349*, 2022.

[33] M. Wermelinger, R. Johns, F. Gramazio, M. Kohler, and M. Hutter, "Grasping and object reorientation for autonomous construction of stone structures," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5105–5112, 2021.

[34] Y. Dai, M. Yang, B. Dai, H. Dai, O. Nachum, J. Tenenbaum, D. Schuurmans, and P. Abbeel, "Learning universal policies via text-guided video generation," *arXiv preprint arXiv:2302.00111*, 2023.

[35] T. Yu, T. Xiao, A. Stone, J. Tompson, A. Brohan, S. Wang, J. Singh, C. Tan, J. Peralta, B. Ichter, *et al.*, "Scaling robot learning with semantically imagined experience," *arXiv preprint arXiv:2302.11550*, 2023.

[36] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[37] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[38] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[39] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning." http://pybullet.org, 2016–2021.

[40] I. A. Şucan, M. Moll, and L. E. Kavraki, "The Open Motion Planning Library," *IEEE Robotics & Automation Magazine*, vol. 19, pp. 72–82, December 2012. https://ompl.kavrakilab.org.

[41] Q. Zhang and Y. Chen, "Fast sampling of diffusion models with exponential integrator," *arXiv preprint arXiv:2204.13902*, 2022.