ReorientDiff: Diffusion Model based Reorientation for Object Manipulation

Anonymous

Abstract— **The ability to manipulate objects in 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 95.2% in simulation. Overall, we present a promising approach to address the reorientation challenge in manipulation by learning a conditional distribution, which is an effective way to move towards generalizable object manipulation. More results can be found on our website: https://sites.google.com/view/reorientdiff/.**

I. Introduction

Rearranging objects into specific poses is a fundamental task. It's not only essential for everyday activities at home but also plays a critical role in industrial applications like packing and assembly lines. Performing such a task requires extracting object information from visual-sensor data and planning a pick-place sequence [2], [3]. 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 an effective strategy when successfully changing an object's pose allows its placement at the target pose [1]. Such a strategy ensures feasible intermediate transition poses in scenarios without common grasps between the current pose and an object's desired placement pose.

The problem of finding reorientation poses is traditionally approached via rejection sampling based on finding successful grasps between the current pose-intermediate pose and intermediate pose-target pose. While previous classical approaches achieve this 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, there have been efforts to improve the reorientation process via a data-driven rejection sampling solution using learned models [1] that predict the feasibility score of an intermediate pose w.r.t. feasible grasps in the current and target pose. While their method improves the success rate and planning time, the algorithm requires processing significantly large number of candidate random samples and specifying the target object's placement pose. The former limits *scalability*, and the latter challenges *generalizability*. Lately, with the advances in language descriptor foundation models like CLIP [5], which projects images and texts to a common feature space, target object 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 introduce ReorientDiff, a diffusion model based *generative* method to restructure the reorientation pose generation pipeline as a conditional distribution learning problem. Such a method enables us to directly sample feasible reorientation poses without rejection sampling, thus improving *scalability*. Our contributions can be summarized as follows:

Learning a distribution of intermediate poses: For a given pile of objects, a target object, and its target placement location, we formulate a conditional distribution of feasible intermediate poses. As compared to rejection sampling using random prior, our approach aims at providing a learned prior to efficiently sample high-quality reorientation poses. Leveraging the multi-modality of diffusion models, this distribution encompasses all poses reachable from both the current pose and the target pose.

Flexibly sampling based on possible grasp poses: It is necessary to make sure that the grasp poses w.r.t. object is constant during one pick-place transition. To achieve this, we flexibly sample intermediate poses from the learned distribution based on feasible grasp poses using classifier guidance via pre-trained success classifiers [7], [1]. Such models implicitly refine sampled pose and operate individually for both transitions during reorientation. Hence, the learned distribution can be used for any possible grasp pose based on kino-dynamic feasibility directly at inference.

Representing target placement location via natural language: We leverage CLIP [5] to generate information embeddings from visual input and task descriptions in natural language. We further use these embeddings as conditions for learning the conditional distribution. While this has been explored in recent literature [6], we see this as a substantial improvement over the baseline.

In the proposed approach, we combine a generic classifier-

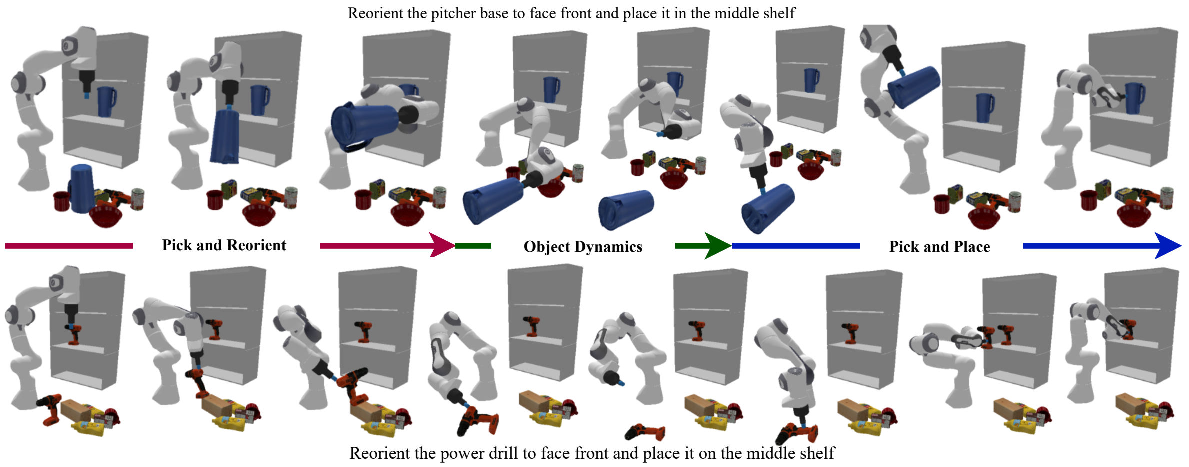

Reorient the pitcher base to face front and place it in the middle shelf
Pick and Reorient
Object Dynamics
Pick and Place
Reorient the power drill to face front and place it on the middle shelf

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 placed 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.* [1], by which we ensure that un-grasping an object in an unstable pose will eventually allow the object to settle at some favourable pose.

free conditional sampling [8] with classifier-guided sampling [9] to sample from diffusion models. To validate the performance of ReorientDiff, we consider reorientation of objects in the YCB dataset [10] that are feasible for suction grippers. For each selected object, we choose suitable locations on multiple shelf levels and target orientations.

II. Related Work

Object manipulation and reorientation. Finding the grasp pose that is feasible for both the current and target location is a widely employed strategy for pick-and-place operations [11], [12], [13]. Such problems are usually solved in two steps: deciding an appropriate placement pose (within a region of interest) and searching for common grasps. In order to ensure feasible target placement, prior works have mostly relied on known object geometries [11], [12], vision-based object representation [2], [14] or using segmentation and depth maps of the pre-specified target object [15], [16], [1]. These strategies have led to several object rearrangement methods [6], [17], [3]. Unlike most prior works that consider the availability of common grasps by default, for complex manipulation scenarios where there are no common grasps, *reorientation* becomes mandatory. The object needs to be reoriented to an intermediate pose and *regrasped* to place it at the target location. Such a scenario has been traditionally tackled via rejection sampling strategies and recently improved via regression-based methods. We also aim to develop a learning-based method.

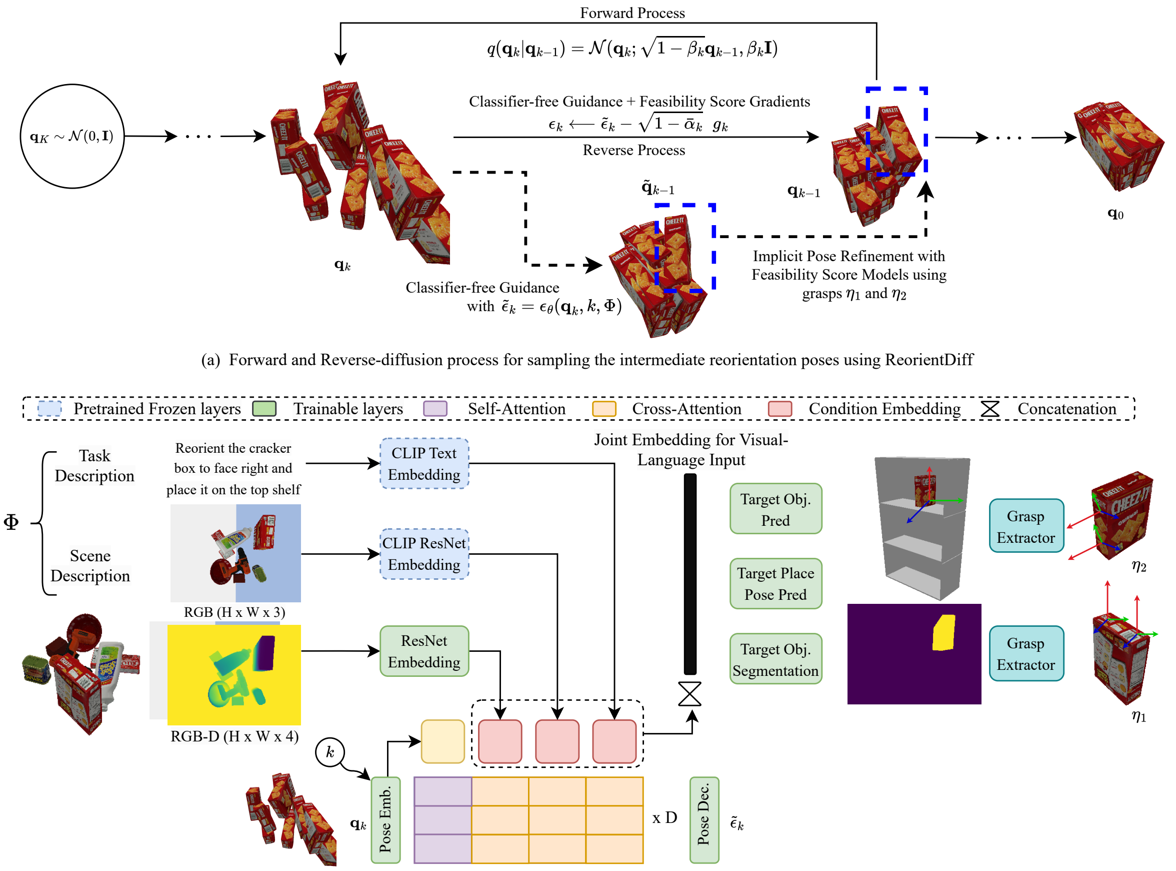
Learning for object manipulation. While prior works have pre-dominantly incorporated trajectory planners [4], they have employed learning strategies to decide the target object and its placement pose as discussed in the previous subsection. Additionally, task descriptions as natural language have been very effective for generalized pick-place tasks in planar tabletop [6] and 3D [17] manipulation. Such language descriptions can be embedded into the learning pipeline via foundation models like CLIP [5], which encodes visual and language information into a common representation space. This has been further extended towards language-conditioned object rearrangement planning [18], [19] and supplying high-level instructions for long-horizon planning [20].

Recently, *reorientation* problems have been solved by planning to reorient objects using extrinsic supports [21], [22], which enables them to re-grasp the object in a desired way. The above methods are regression-based and limited to modeling only one solution pose. Such approaches cannot cater to the multiple possible solutions of the same problem. In such a case, rejection sampling is still beneficial and can be performed using learned feasibility prediction models [1]. We want to develop a pipeline that can still learn about all feasible poses without analyzing extensive random samples.

Generative models for object manipulation. For pick-and-place and reorientation tasks, there can be multiple feasible grasps and reorientation poses respectively. Hence, generative models offer an option to learn them as conditional distributions. Prior works have explored VAE for planning grasps [7] using visible point-cloud of objects. In this direction, diffusion models have been shown to be advantageous for robotics [23], [24], [25], [26], [27]. Recent works have demonstrated the multi-modal distribution learning using diffusion models for finding target poses [18], [28] and learning policies [23], [24], [25]. In addition to such properties, we also plan to leverage the flexible sampling and conditioning strategies offered by diffusion models to incorporate additional conditions at inference without re-training.

III. Preliminary: Diffusion Models

Consider samples x_0 from an unknown data distribution $q(x_0)$; diffusion models [29] 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


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 $\tilde{\epsilon}_k = \epsilon_\theta(\mathbf{q}_k, k, \Phi)$
Implicit Pose Refinement with Feasibility Score Models using grasps η_1 and η_2
Pretrained Frozen layers
Trainable layers
Self-Attention
Cross-Attention
Condition Embedding
Concatenation
Φ
Task Description
Reorient the cracker box to face right and place it on the top shelf
Scene Description
RGB (H x W x 3)
RGB-D (H x W x 4)
CLIP Text Embedding
CLIP ResNet Embedding
ResNet Embedding
Joint Embedding for Visual-Language Input
Target Obj. Pred
Target Place Pose Pred
Target Obj. Segmentation
Grasp Extractor
Grasp Extractor
η_2
η_1
k
$\mathbf{q}_k$
Pose Emb.
x D
Pose Dec.
$\tilde{\epsilon}_k$

(a) Forward and Reverse-diffusion process for sampling the intermediate reorientation poses using ReorientDiff

(b) Classifier-free diffusion score function model for $\tilde{\epsilon}_k = \epsilon_\theta(\mathbf{q}_k, k, \Phi)$

Fig. 2: **Method Overview** (a) **Forward and reverse diffusion process.** ReorientDiff uses a combination of classifier-free guidance with classifier-based implicit refinement to sample from the learned distribution of intermediate poses. It ensures high-success feasibility with minimal variance by guiding the scene-task conditioned sampling using feasibility score gradients. (b) **Conditioned score function.** ReorientDiff learns the target distribution of feasible reorientation poses conditioned on the scene (pile of objects) and task (language prompt) jointly represented as Φ. We use the pre-trained frozen CLIP text and image embeddings to formulate a joint embedding, trained end-to-end to encode information about placement pose, target object and current pose. Further, the current pose and target poses are processed to obtain feasible grasps (η_1 and η_2), which are used to calculate the feasibility gradients g_k in (a). The joint embedding is used as a sequence to condition the transformer-based score network $\epsilon_\theta(\mathbf{q}_k, k, \Phi)$ via cross-attention to obtain the classifier-free score estimate in (a).

in x_0 to create a Markov chain with latents $x_{1:K}$ following transitions: $q(x_{1:K}|x_0) = \prod_{k=1}^{K} q(x_k|x_{k-1})$, 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 $\beta_1, \ldots, \beta_K$. 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 [30] 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 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

$$\mu_\theta(x_k, k) = \frac{1}{\sqrt{\alpha_k}}\left(x_k - \frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}}\epsilon_\theta(x_k, k)\right). \qquad (1)$$

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 [31]

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

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

$$\nabla_{x_k}\log p_\theta(x_k) = -\frac{1}{\sqrt{1-\bar{\alpha}_k}}\epsilon_\theta(x_k, k). \qquad (3)$$

IV. Reorientation

Reorientation consists of solving two problems simultaneously, finding a pose that is reachable from the current pose and, after the effect of gravity, results in a pose that makes placement at the target pose achievable (as shown in Figure 1). Once we have an estimate of the current and target pose, it is intuitive that there will be a set of poses that will satisfy reorientability. However, only a small subset of such reorientable poses will be valid with provided kinodynamic constraints on grasp poses. Identifying a candidate sample from this subset by either brute force sampling or optimization is computationally expensive and has to be done for every new scenario.

To circumvent the above challenges, we propose a generative modeling approach to sample from the subset of valid reorientation poses. More specifically, our method learns the distribution of all reorientable poses using a conditional diffusion model and use classifiers to guide sampling towards valid poses directly during inference based on provided grasp poses. Hence, we divide the problem into three segments: i) regression-based end-to-end learning for finding the target object and placement pose from the scene and task description (scenario), ii) learning the distribution of all reorientable poses for a given scenario once the object specifications are known and iii) learning grasp feasibility classifiers for selecting only the valid reorientation poses. To achieve this, we discuss our formulation for constructing scene-task representation, calculating grasp poses from object poses and learning grasp feasibility classifiers below. The diffusion model training and inference is discussed in the next section.

A. Constructing Generic Scene-Task Representations

A scene-task representation is a compact embedding of all available information present in the scene and specified by the user. 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. For learning the semantic and spatial embeddings [6], [17], we use pre-trained CLIP foundation model and obtain semantic embeddings from the image $\mathcal{I}$ and language $\mathcal{L}$. We sequence the embeddings with the spatial embeddings for target object segmentation to get a joint embedding sequence Φ as generic scene-task representation as shown in Figure 2(b). The embedding is further used to predict the target object 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 [1]. The predicted target object segmentation 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. Feasibility Score Models

Following prior works [7], [1], [19], a feasibility prediction model is important for early-evaluation and rejection of unfavorable 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 [1], is particularly interesting 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 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 the 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

$$p_h(\mathbf{q}) \propto p(\mathbf{q}|\Phi) h(\mathbf{q}, \Phi), \tag{4}$$

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 (4), we use classifier-free guidance [8] to sample high-likelihood reorientation poses for a particular scene-task representation. We train a score-network [31], $\epsilon_\theta(\mathbf{q}_k, \Phi) \propto \nabla_{\mathbf{q}_k} \log p(\mathbf{q}_k|\Phi)$, 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

$$\tilde{\epsilon}_k = \epsilon_\theta(\mathbf{q}_k, \Phi) + w_c \Big(\epsilon_\theta(\mathbf{q}_k, \Phi) - \epsilon_\theta(\mathbf{q}_k, \varnothing)\Big) \tag{5}$$

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 [30] sampling strategy and $\tilde{\epsilon}_k$ as follows:

$$\tilde{\mathbf{q}}_{k-1} \longleftarrow \sqrt{\bar{\alpha}_{k-1}} \Big(\frac{\mathbf{q}_k - \sqrt{1-\bar{\alpha}_k}\ \tilde{\epsilon}_k}{\sqrt{\bar{\alpha}_k}} \Big) + \sqrt{1-\bar{\alpha}_{k-1}}\ \tilde{\epsilon}_k \tag{6}$$

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 [9] formulation for the heuristics, the reverse diffusion can be formulated as:

$$\begin{aligned} & p_h(\mathbf{q}_k|\mathbf{q}_{k+1}, y, \Phi) \propto \\ & p(\mathbf{q}_k|\mathbf{q}_{k+1}, \Phi)\ \ p(y|\eta_1, \hat{\mathbf{q}}_0^k, \Phi)|_{\mathcal{M}_1}\ p(y|\eta_2, \hat{\mathbf{q}}_0^k, \Phi)|_{\mathcal{M}_2} \end{aligned} \tag{7}$$

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

$$\hat{\mathbf{q}}_0^k = \frac{\mathbf{q}_k - \sqrt{1-\bar{\alpha}_k}\ \ \tilde{\epsilon}_k}{\sqrt{\bar{\alpha}_k}} \tag{8}$$

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 (7) as:

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

In view of (1), we then obtain the modified score

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

where $g_k = -\beta_k \sum_{i=1}^{2} w_i \nabla_{\mathbf{q}_k} \Big[1 - \mathcal{M}_i(y|\eta_i, \hat{\mathbf{q}}_0^k, \Phi)\Big]^2$. 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 (6). A visual clarification of the forward and reverse diffusion is shown in Figure 2(a).

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.

A. Dataset Generation and Training

We use PyBullet [32] and an OMPL [33] 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 converted goal poses into modular language instructions, and the success of pick and place for both the steps was recorded for 10000 scenarios. 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 when combined with the feasibility score models.

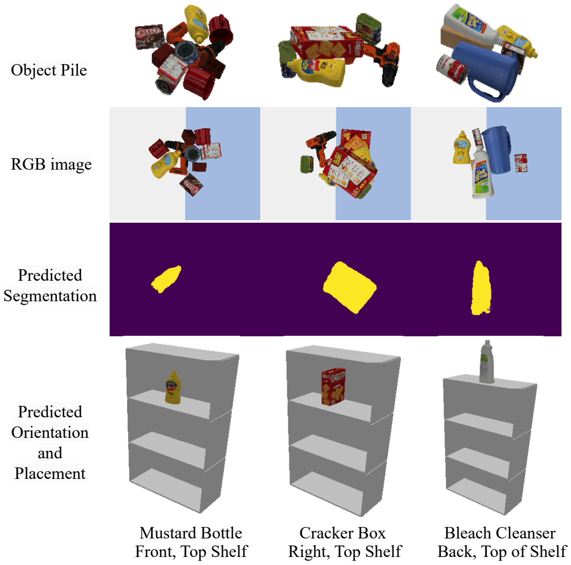

Object Pile
RGB image
Predicted Segmentation
Predicted Orientation and Placement
Mustard Bottle Front, Top Shelf
Cracker Box Right, Top Shelf
Bleach Cleanser Back, Top of Shelf

Fig. 3: **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).

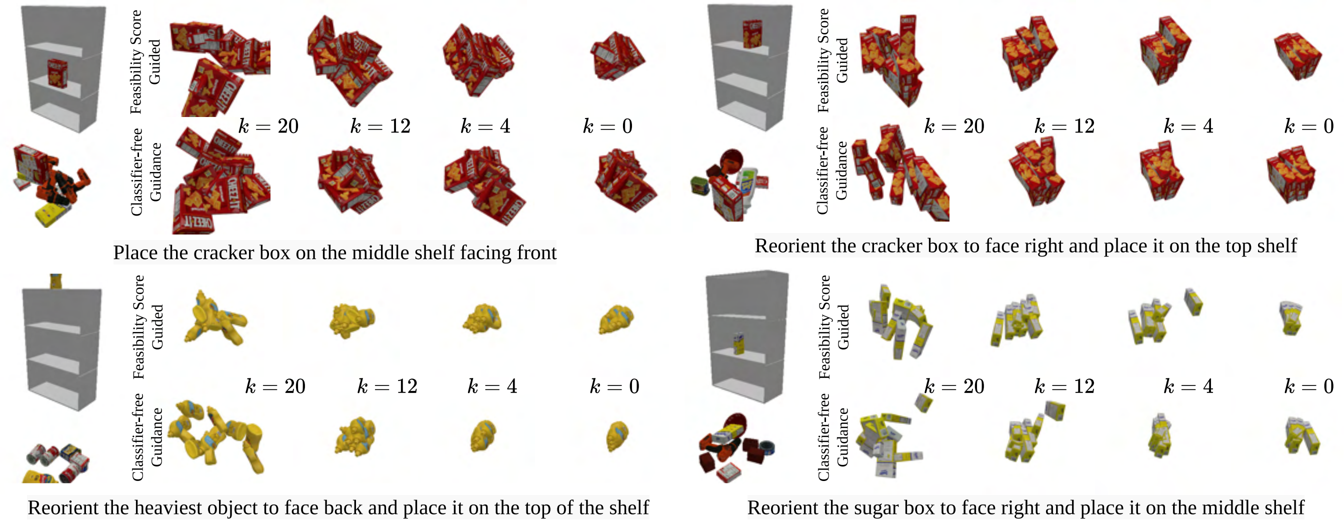
B. Performance Evaluation: Scene-Task Representation

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 Figure 3 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. In our experiments, 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 Figure 4 for a set of YCB Objects [10] and different scene-task scenarios where only 40 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 [1] in Table I.

The i) reorientation success measures the capability of the diffusion model to generalize to poses which ensure reorientability, ii) placement success measures the successful



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. 4: **Reverse Diffusion for Reorientation Pose Generation** The reverse sampling process for 4 k-steps at $k = 20, 12, 4, 0$ for $K = 20$ 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.

TABLE I: Success evaluation of the proposed method as compared to the rejection sampling based baseline ReorientBot. The ReorientDiff algorithm was tested for more than 100 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 3 random seeds.

Method	Success (%) Reorient	Success (%) Place	Success (%) Overall
Random	43.4	40.8	40
ReorientBot	97.9	95.1	93.2
ReorientDiff (w/o Guide)	97.4	92.3	90.8
ReorientDiff	**98.9**	**96.5**	**95.2**

placement of object in the and iii) overall success indicates that the desired pose is achieved while 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.

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

ReorientDiff K	Success (%) Reorient	Success (%) Place	Success (%) Overall
$K = 10$	97.4	94.5	93.9
$K = 20$	**98.9**	**96.5**	**95.2**

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. ReorientDiff performs well with only 20 sampling steps.

Following our analysis on performance, we explored the time consumption for the 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 and the baseline in Table III.

TABLE III: Computational analysis of the planning time for ReorientDiff ($K = 20$) with and without feasibility score guidance along with the baseline.

Method	Planning Time (sec)
ReorientBot	2.5
ReorientDiff (w/o Guide)	0.3
ReorientDiff	**1.05**

Our findings show that ReorientDiff leverages fast sampling strategies of FastDPM [34] to recover from computationally heavy gradient calculations for reverse denoising steps. Without using the guidance from the feasibility-score models, classifier-free guidance requires even less time as compared to the baseline, ReorientBot, as shown in Table III. Hence, from our visual and empirical analysis, ReorientDiff successfully proves that formulating the problem of reorientation as learning a conditional distribution is an efficient and scalable way to move towards more generalizable object manipulation.

VII. Conclusion

Diffusion models are powerful generative models capable of modeling (conditional) distributions. Our 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 the success of the planned intermediate poses. With only 10 candidate reorientation poses, we achieved an overall success rate of 95.2% across various objects. With the possible inclusion of point-cloud-based object representations [28], such a method can generalize to a more diverse set of objects.

References

[1] 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.

[2] 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.

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

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

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

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

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

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

[24] 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.

[25] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *arXiv preprint arXiv:2303.04137*, 2023.

[26] U. A. Mishra, S. Xue, Y. Chen, and D. Xu, "Generative skill chaining: Long-horizon skill planning with diffusion models," in *7th Annual Conference on Robot Learning*, 2023.

[27] Z. Xian, N. Gkanatsios, T. Gervet, and K. Fragkiadaki, "Unifying diffusion models with action detection transformers for multi-task robotic manipulation," in *7th Annual Conference on Robot Learning*, 2023.

[28] A. Simeonov, A. Goyal, L. Manuelli, L. Yen-Chen, A. Sarmiento, A. Rodriguez, P. Agrawal, and D. Fox, "Shelving, stacking, hanging: Relational pose diffusion for multi-modal rearrangement," *arXiv preprint arXiv:2307.04751*, 2023.

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

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

[31] 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.

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

[33] 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`.

[34] Z. Kong and W. Ping, "On fast sampling of diffusion probabilistic models," *arXiv preprint arXiv:2106.00132*, 2021.