DALL-E-Bot: Introducing Web-Scale Diffusion Models to Robotics

Ivan Kapelyukh*[1,2], Vitalis Vosylius*[1], Edward Johns[1]

***Abstract*— We introduce the first work to explore web-scale diffusion models for robotics. DALL-E-Bot enables a robot to rearrange objects in a scene, by first inferring a text description of those objects, then generating an image representing a natural, human-like arrangement of those objects, and finally physically arranging the objects according to that image. The significance is that we achieve this zero-shot using DALL-E, without needing any further data collection or training. Encouraging real-world results with human studies show that this is a promising direction for the future of web-scale robot learning. We also propose a list of recommendations to the text-to-image community, to align further developments of these models with applications to robotics. Videos are available on our webpage at: https://www.robot-learning.uk/dall-e-bot**

I. Introduction

Web-scale image diffusion models, such as OpenAI's DALL-E 2 [1], have been one of the most exciting recent breakthroughs in machine learning. By training over hundreds of millions of image-caption pairs from the Web, these models learn a language-conditioned distribution over natural images, from which novel images can be generated given a text prompt. Large language models [2], [3], which are also trained on web-scale data, have recently been explored for robotics applications [4], [5] to enable generalisation of language-conditioned policies to novel language commands. Given these successes, in this paper we ask the following question: **Can web-scale text-to-image diffusion models, such as DALL-E, be exploited for real-world robotics?**

Since these models can generate realistic images of everyday scenes such as kitchens and offices, our insight is that they are proficient at imagining arrangements of everyday objects which are *human-like*: semantically correct, aesthetically pleasing, physically plausible, and convenient to use. Therefore, we propose that they could be used to generate images of goal states for generic object rearrangement tasks [6], such as setting a table, loading a dishwasher, tidying a room, stacking a shelf, and assembling furniture.

In this paper, we propose DALL-E-Bot, the first method to explore web-scale image diffusion models for robotics. We design a framework which enables DALL-E to be used to predict a goal state for object rearrangement, given an image of an initial, disorganised scene, as shown in Fig. 1. The pipeline converts the initial image into a text caption, which is then passed into DALL-E to generate a new image, from which we then obtain goal poses for each object. Note that we use publicly-available DALL-E as it is, without requiring any further data collection or training. This is important: it allows for zero-shot, open-set, and autonomous rearrangement, going beyond prior work which often requires collecting examples of desirable arrangements and training a model specifically for those scenes [7], [8], [9], [10].

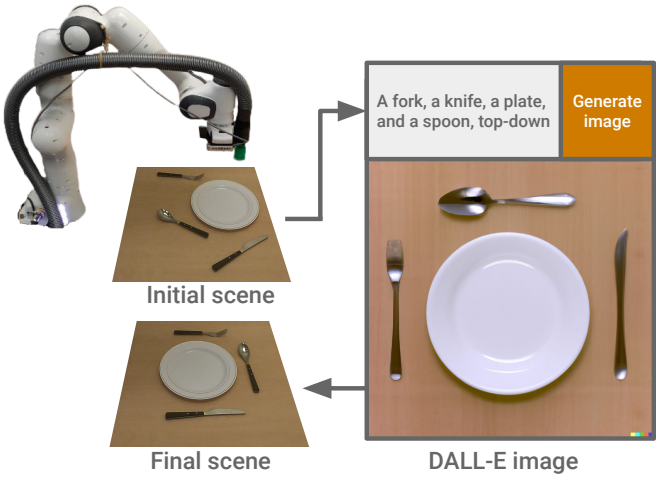

Fig. 1: In DALL-E-Bot, the robot prompts DALL-E with a list of objects it has detected, which then generates an image of a human-like arrangement of those objects. The robot then rearranges the objects via pick-and-place to match the generated image.

With real-world experiments across three everyday rearrangement tasks, we studied two abilities of DALL-E-Bot: (1) to re-arrange a full scene by generating an entire image, and (2) to precisely place a single item into an existing scene using DALL-E's inpainting feature. By evaluating with human users, we found that DALL-E-Bot arranges scenes in an appealing way to humans, and we also explored the impact of our design choices with ablation studies. Finally, we present the limitations of current web-scale diffusion models for robotics, and propose recommendations to the text-to-image community for better aligning these models with robotics applications.

II. Related Work

A. Predicting Goal Arrangements

We now highlight prior approaches to predicting goal poses for rearrangement tasks. Some methods view the prediction of goal poses as a classification problem, by choosing from a set of discrete options for an object's placement. For house-scale rearrangement, a pre-trained language model can be used to predict goal receptacles such as tables [11], and out-of-place objects can be detected automatically [12]. At a room level, the correct drawer or shelf can be classified [13], taking preferences into account [14]. Lower-level prediction

* Joint first authorship. [1] The Robot Learning Lab at Imperial College London. [2] The Dyson Robotics Lab at Imperial College London.

from a dense set of goal poses can be achieved with a graph neural network [15] or a preference-aware transformer [10]. Our framework uses high-resolution images of how objects should be placed, thus not requiring a set of discrete options to be pre-defined, and predicting more precise poses than is possible with language. Prior robotics work has trained generative models for visual control [16], [17], but our work shows that web-scale models such as DALL-E can be used zero-shot, even for multi-stage rearrangement tasks.

Methods for predicting continuous object poses typically use a dataset of example arrangements. They can learn spatial preferences with a graph VAE [8], or model gradient fields [18]. For language-conditioned rearrangement, an autoregressive transformer [9] can be used, or a diffusion model over poses can be combined with learned discriminators [19]. Other methods use full demonstrations [20], [21], or leverage priors such as human pose context [7]. However, unlike these works, our proposed framework does not require collecting and training on a dataset of rearrangement examples, which often restricts these methods to a specific set of objects and scenes. Instead, exploiting existing web-scale image diffusion models enables zero-shot rearrangement. When the goal image is given, rearrangement is possible even with unknown objects [22]. Our method does not require a user-provided goal image, and is thus an autonomous system.

B. Web-Scale Diffusion Models

Generating images with web-scale diffusion models such as DALL-E is at the heart of our method. Diffusion models [23] are trained to reverse a single step of added noise to a data sample. By starting from random noise and iteratively running many of these small, learned denoising steps, this can generate a sample from the learned distribution of data. These models have been used to generate images [24], [25], [26], text-conditioned images [1], [27], [28], [29], robot trajectories [30], and audio waves [31]. We use DALL-E 2 [1] in this work, although our framework could be used with other text-to-image models.

III. Method

A. Overview

We address the problem of predicting a goal pose for each object in a scene, such that objects can then be rearranged in an appealing way. We propose to predict goal poses zero-shot from a single RGB image I_I of the initial scene.

To achieve this, we propose a modular pipeline shown in Fig. 2. At the heart of our method is a web-scale image diffusion model DALL-E 2 [1], which, given a text description ℓ of the objects in a scene, can generate a goal image I_G, depicting a human-like arrangement of those objects. We can sample many such images for a given text description. We convert an initial RGB observation into a more relevant object-level representation to individually reason about the objects in the scene. This representation consists of text captions of crops of individual objects c_i (used to construct a text prompt ℓ) together with their segmentation masks M_i, and visual-semantic feature vector v_i acquired using the CLIP model [32]. We also convert generated images into object-level representations and select the image that has the same number of objects as the initial scene, and best matches the objects in the initial scene semantically. Using an Iterative Closest Point (ICP) [33] algorithm in image space, we then register corresponding segmentation masks to obtain transformations that need to be applied to the individual objects to achieve the desired arrangement. Finally, we convert these transformations from image space to Cartesian space using a depth camera observation, and deploy a real Franka Emika Panda robot equipped with a compliant suction gripper to re-arrange the scene. Since this method is modular, it will improve as the individual components (e.g. segmentation) improve in the future.

B. Object-Level Representation

To reason about the poses of individual objects in the observed scene, we need to convert the initial RGB observation into a more functional, object-level representation. We use the Mask R-CNN model [34] from the Detectron2 library [35] to detect objects in an image and generate segmentation masks M_i. This model was pre-trained on the LVIS dataset [36], which has 1200 object classes, being more than sufficient for many rearrangement tasks. For each object, Mask R-CNN provides us with a bounding box, a segmentation mask, and a class label. However, we found that whilst the bounding box and segmentation mask predictions are usually high quality and can be used for pose estimation (described in Section III-E), the predicted class labels are often incorrect due to the large number of classes in the training dataset.

As we are using labels of objects in the scene (described in Section III-C) to construct a prompt for an image diffusion model, it is crucial for these labels to be accurate. Therefore, instead of directly using predicted object class labels, we pass RGB crops around each object's bounding box through an OFA image-to-text captioning model [37], and acquire text descriptions of the objects in the initial scene observation, c_i. Generally, this approach allows us to more accurately predict object class labels and go beyond the objects in Mask R-CNN's training distribution, and even obtain their visual characteristics such as colour or shape. Finally, we also pass each object crop through a CLIP visual model [32], giving each object a 512-dimensional visual-semantic feature vector v_i. These features will be used later for matching objects between the initial scene image and the generated image.

In summary, by the end of this stage, we have converted an RGB observation I_I into an object-level representation (M_i, c_i, v_i), which represents each object by a segmentation mask, a text caption, and a semantic feature vector.

C. Goal Image Generation

Our method relies on the ability to generate images of natural and human-like arrangements, given their text descriptions. To this end, we exploit recent advances in text-to-image generation and web-scale diffusion models, by using the publicly-available DALL-E 2 [1] model from

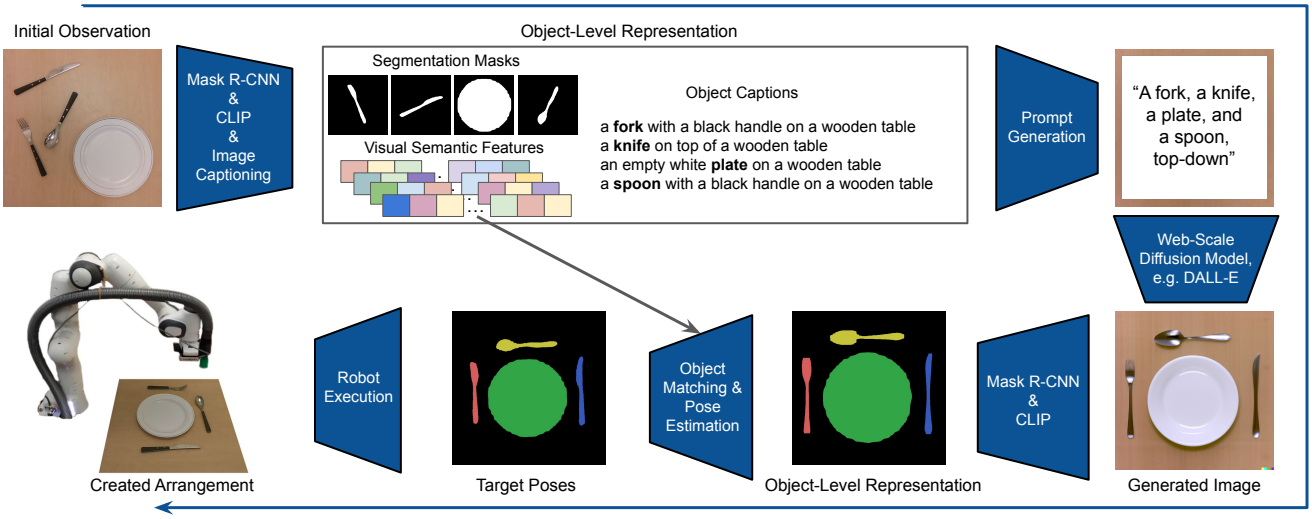

Fig. 2: DALL-E-Bot creates a human-like arrangement of objects in the scene using a modular approach. First, the initial observation image is converted into a per-object description consisting of a segmentation mask, an object caption, and a CLIP visual feature vector. Next, a text prompt is constructed describing the objects in the scene and is passed into DALL-E to create a goal image for the rearrangement task, where the objects are arranged in a human-like way. Then, the objects in the initial and generated images are matched using their CLIP visual features, and their poses are estimated by aligning their segmentation masks. Finally, a robot rearranges the scene based on the estimated poses to create the generated arrangement.

OpenAI. This has been trained on a vast number of image-caption pairs from the Web, and represents the conditional distribution $p_\theta(I_G|\ell, I_M)$. Here, I_G is an image generated by the model, ℓ is a text prompt, and I_M is an image mask that can be used to prevent the model from changing the values of certain pixels in the image. A large portion of distribution p_θ represents images with scenes arranged by humans in a natural and usable way. Therefore, by sampling from this distribution, we can generate images depicting human-like arrangements and create those arrangements in the real world by moving objects to the same poses as in the generated images. Additionally, the ability to condition this distribution on an image mask I_M lets us handle cases where not all objects in the scene can (or should) be moved by the robot.

To generate an image using DALL-E, we first need to construct a text prompt ℓ describing the desired scene. To this end, we use object captions from our object-level representation. Although full captions, including visual characteristics, could be used to generate images with objects closely resembling the observed ones, in this work, we only use the nouns describing the object's class and leave including visual characteristics for future work. Thus, we extract the class of each object from the caption of its object crop, i.e. we extract "apple" from "a red apple on a wooden table". We do this by passing the object captions through the Part-of-Speech tagging model [38] from the Flair NLP library [39], which tags each word as a noun, a verb, etc. From this list of classes, we construct a prompt that makes minimal assumptions about the scene, to allow DALL-E to arrange it in the most natural way. In this work, our experiments are based on tabletop scenes, with observations captured by a camera mounted on a robot's wrist pointing downwards towards the table. Therefore, we added a "top-down" phrase to the prompt to better align the initial and generated images. As such, an example prompt we use would be "A fork, a knife, a plate, and a spoon, top-down" (as in Fig. 2).

We use DALL-E's ability to condition distribution p_θ on image masks in three ways. First, if there are objects in the scene that a robot is not allowed to move, we add their contours to I_M. This prevents DALL-E from generating these objects in different poses while still allowing for other objects to be placed on top or in them (e.g. a basket can not be moved, but other objects can be placed inside it). Second, we add a mask of the tabletop's edges in our scene to I_M to visually ground the generated images. This prevents objects from being placed on the edge of the generated image. Additionally, we found throughout our experiments that this incentivises DALL-E to create objects of appropriate sizes. Third, we subtract segmentation masks of all the movable objects from I_M, with enlarged masks to remove any shadows. Avoiding shadows is essential, as if DALL-E sees any shadows of objects in their original poses, it will generate objects in the same poses to fit with those shadows, hindering its ability to generate novel and diverse images.

Using the prompt ℓ and the image mask I_M, we sample a batch of images from the conditional distribution $p_\theta(I_G|\ell, I_M)$, representing the text-to-image model. We do so using an automated script and OpenAI's web API.

D. Image Selection & Object Matching

In the batch of four images generated by DALL-E, not all will be desirable for the rearrangement task; some may have artefacts hindering object detection, others may include extra objects that were not part of the text prompt, etc. Therefore,

we need to select the generated image I_G whose objects best match those in the real-world initial image I_I.

For each generated image, we obtain segmentation masks and a CLIP semantic feature vector for each object, using the procedure in Section III-B. Then, we filter out generated images where the number of objects is different to the initial scene. If there are no images with the same number of objects in the DALL-E-generated images, we sample another batch. We then match objects between the generated image and initial image. This is non-trivial since the generated objects are different instances to the real objects, often with a very different appearance. Inspired by [40], we compute a similarity score between any two objects (one from I_I, and one from I_G) using the cosine similarity between their CLIP feature vectors. Since greedy matching is not guaranteed to yield optimal results in general, we use the Hungarian Matching algorithm [41] to compute an assignment of each object in the initial image to an object in the generated image, such that the total similarity score is maximised. Then, we select the generated image I_G which has the best overall score with the initial image I_I. This image depicts the most similar set of objects to the real scene, and therefore gives the best opportunity for rearranging the real scene.

E. Object Pose Estimation

For each object in the initial image, we now know its segmentation mask in the initial image and the corresponding segmentation mask in the generated image. By aligning these masks, we can estimate a transformation from the initial pose (in the initial image) to the goal pose (in the generated image). We rescale each initial segmentation mask, such that the dimensions of its bounding box equal those in the generated image, and then use the Iterative Closest Point (ICP) algorithm [33] to align the two masks, taking each pixel to be a point. This gives us a 3-DoF (x, y, θ) transformation $\mathcal{T}$ in pixel space between the initial and goal pose. We run ICP from many random initial poses, due to local optima. For objects with nearly symmetric binary masks such as knives, aligning masks with ICP leads to multiple candidate solutions (for knives, they differ by 180 degrees). To select the correct solution (handle aligned with handle, blade aligned with blade), we pass the generated object image o_G and the transformed real object image $\mathcal{T}(o_I)$ through a semantic feature map extractor f_S (an ImageNet-trained ResNet [42], [43]). We select the ICP solution $\mathcal{T}$ which minimises the photometric loss between the semantic feature maps: $\mathcal{L}_S = (f_S(o_G) - f_S(\mathcal{T}(o_I)))^2$.

The generated image can depict objects of a different scale than the objects in the initial image. Naively moving objects to estimated poses can lead to collisions (if generated objects are smaller) or unnaturally spaced-out arrangements (if generated objects are larger). Therefore, we move objects closer together or further apart based on the mismatch in size. Additionally, we ensure there are no collisions in the arrangement by moving any colliding objects further apart.

Next, we use a wrist-mounted depth camera to project the pixel-space poses into 3D space on the tabletop, to obtain a transformation for each object which would move it from the initial real-world pose to the goal real-world pose. Finally, the robot executes these transformations by performing a sequence of pick-and-place operations using a suction gripper. We also designed a simple planner which first moves objects that would cause collisions into intermediate poses to the side, before later moving them to their goal poses.

IV. Experiments

In our experiments, we evaluate the ability of our method to create human-like arrangements using both subjective (Section IV-A) and objective (Section IV-B) metrics.

A. Zero-Shot Autonomous Rearrangement

First, we explore the following question: **Can DALL-E-Bot arrange a set of objects in a human-preferred way?** We evaluate on 3 everyday tabletop rearrangement tasks: *dining scene*, *office scene*, and *fruit scene* (Fig. 3), where the robot should arrange the objects in a human-like way in each scene. The dining scene contains four objects: a knife, a fork, a spoon, and a plate. The office scene contains a stationary iPad which the robot can see but is not allowed to move, and three movable objects: a keyboard, a mouse, and a mug. The fruit scene contains a stationary basket, and three movable objects: two apples and an orange.

In real-world applications of DALL-E-Bot, users would only see the outcome of the real-world rearrangement, which would include all the errors that might accumulate through the pipeline. To simulate this experience for our evaluation, we create the predicted arrangements using a real-world Franka Emika robot equipped with a compliant suction gripper. Then, we record the outcome as an RGB image of a tabletop scene, using a camera on the wrist of the robot.

Since DALL-E-Bot is the first method to predict precise goal poses for rearrangement in a way which is zero-shot and autonomous, we designed two baselines which are also zero-shot for a fair comparison. The *Rand-No-Coll* baseline places objects randomly in the environment while ensuring they do not overlap. The *Geometric* baseline puts all the objects evenly in a straight line such that they are not colliding, and aligns the objects so that they are parallel using their bounding boxes. In addition, we compare our method to two different variants. *DALL-E-Bot-AR* creates an arrangement in an auto-regressive way, with a sequence of goal images rather than a single image, where each placed object is treated as a stationary object for the next generated image (and thus its contours are added to I_M). Here, we do not adjust the poses of the objects based on the size mismatch and do not reject generated images with the wrong number of objects. Finally, *DALL-E-Bot-NF* (no filtering) does not filter generated images and always uses the first image. If this image has fewer objects than in the real scene, unmatched objects are placed randomly, whilst avoiding collisions.

Since we aim to create arrangements which are appealing to humans, the most direct evaluation is to ask humans for feedback. Therefore, we showed human users images of the final real-world scene created by the robot, and

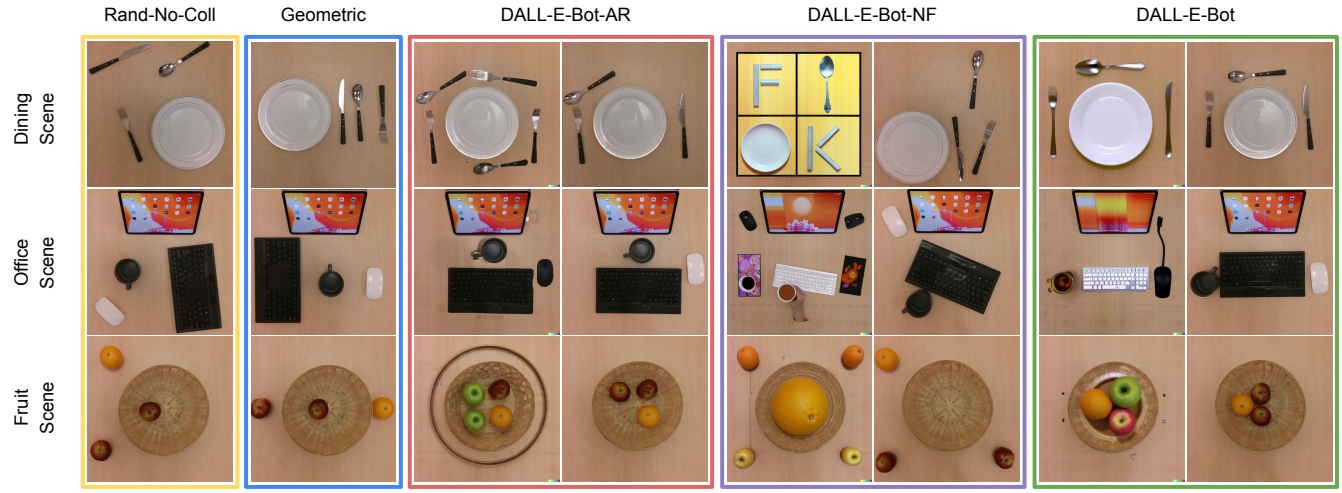

Fig. 3: Examples of scenes rearranged by the robot using different methods. Columns for the methods that use DALL-E include the generated image (left) and the final arrangement (right). For *DALL-E-Bot-AR*, images are from the last step.

asked them the following question: *"If the robot made this arrangement for you at home, how happy would you be?"*. The user provided a score for each method on a Likert Scale from 1 (very unhappy) to 10 (very happy), while being shown arrangements made by each method side-by-side in a web-based questionnaire. We recruited 40 users representing 18 nationalities, both male and female, with ages ranging from 22 to 71. Each rated the results of 5 methods on 5 random initialisations of 3 scenes, for a total of 3000 ratings. Initialisations were roughly matched for all the methods and all users were shown the same images.

Method	Dining Scene	Office Scene	Fruit Scene	Mean
Rand-No-Coll	2.03±1.34	3.56±2.01	2.94±2.01	2.84
Geometric	4.08±2.27	3.36±2.01	3.13±1.82	3.52
DALL-E-Bot-NF	3.87±2.78	6.54±2.34	7.45±3.19	5.95
DALL-E-Bot-AR	4.88±2.61	7.37±2.05	9.59±0.90	7.28
DALL-E-Bot	**8.01**±2.03	**7.56**±2.02	**9.81**±0.52	**8.46**

TABLE I: User ratings for arrangements by each method. Each cell shows the mean and standard deviation across all users and scene initialisations, with the best in bold.

The results of this user study are in Table I. Example arrangements are shown in Fig. 3, and videos are available at: **https://www.robot-learning.uk/dall-e-bot**. *DALL-E-Bot* receives high user scores, showing that it can create satisfactory arrangements zero-shot, without requiring task-specific training. It beats the heuristic baselines, showing that users value semantic correctness for arranging scenes beyond simple geometric alignment, which justifies the use of web-scale learning of these semantic rules. This is especially evident in the dining scene, where DALL-E recognises the semantic structure which can be created from those objects. The *DALL-E-Bot-NF* ablation performs the worst out of the DALL-E-Bot variants on all scenes. This justifies our sample-and-filter approach for using these web-scale models, which ensures that the robot can feasibly create the generated arrangement, rather than naively using the first generated image. The *DALL-E-Bot-AR* variant performs well generally but struggles in the dining scene, where the thin cutlery may slip, leading to accumulating error since the method auto-regressively conditions on the objects placed so far. *DALL-E-Bot* avoids this issue by jointly predicting all object poses.

B. Placing Missing Objects with Inpainting

In the next experiment, we use objective metrics to answer the following question: **Can DALL-E-Bot precisely complete an arrangement which was partially made by a human?** For this, we ask DALL-E-Bot to find a suitable pose for an object that has been masked out from a human-made scene, while the other objects are kept fixed. We study this using the dining scene, because it has the most semantically rigid structure, lending itself well to quantitative, objective evaluation. To create these scenes, we recruited ten users (both left and right-handed) and asked them the following: "*Imagine you are sitting down here for dinner. Can you please arrange these objects so that you are happy with the arrangement?*". As there can be multiple suitable poses for any object, for each of the objects we asked the users to provide any alternative poses that they would still be happy with, while keeping other objects in their original poses.

Given the image of the arrangement made by a user, we mask out everything except the fixed objects. This means that DALL-E cannot change the pixels belonging to fixed objects. The method must then predict the pose of the missing object. DALL-E-Bot does this by inpainting the missing object somewhere in the image. For a given user, the predicted pose for the missing object is compared against the actual pose in their arrangement. This is done by aligning two segmentation masks of the missing object, one from the actual scene and one at the predicted pose. Since this is for two poses of exactly the same object instance, we find the alignment is highly accurate and can be used to estimate the error between the actual and predicted pose. From this transformation, we take the orientation and distance errors projected into the

workspace as our metrics. This is repeated for every object individually as the missing object, and across all the users.

We compare our method to two zero-shot heuristic baselines, *Rand-No-Coll* and *Geometric*. *Rand-No-Coll* places the missing object randomly within the bounds of the image, ensuring it does not collide with the fixed objects. *Geometric* first finds a line defined by centroids of segmentation maps of two fixed objects. Then it places the considered object on that line such that it is as close to the fixed objects as possible, does not collide with them, and its orientation is aligned with the orientation of the closest object.

We compare the predicted pose against each of the acceptable poses provided by the user, and report the position and orientation errors from the closest acceptable pose in Table II. The distribution of acceptable poses is multimodal. Therefore, we present the median error across all users, which is less dominated by outliers than the mean and is a more informative representation of the aggregate performance. DALL-E-Bot outperforms the baselines, and is able to accurately place the missing objects for different users. This implies that it is successfully conditioning on the poses of the other objects in the scene using inpainting, and that the human and robot can create an arrangement collaboratively.

	Fork	Plate	Spoon	Knife
Method	cm / deg	cm / deg	cm / deg	cm / deg
Rand-No-Coll	25.85 / 70.32	10.78 / -	27.47 / 42.56	23.51 / 99.32
Geometric	15.59 / 40.57	2.29 / -	23.83 / 86.11	11.58 / **1.47**
DALL-E-Bot	**4.95** / **1.26**	**1.28** / -	**2.13** / **2.72**	**2.1** / 3.27

TABLE II: Position and orientation errors between predicted and user preferred object poses. Each cell shows the median across all users, with the best in bold.

V. Discussion

A. Limitations

Top-down pick-and-place. Our experiments focus on 3-DoF rearrangement, which is sufficient for many everyday tasks. However, future work can extend to 6-DoF object poses with more complex interactions, e.g. to stack shelves. This could draw from recent works on collision-aware manipulation [44] and learning of skills beyond grasping [45].

Overlap between objects. Currently, our method assumes that movable objects cannot overlap, e.g. the fork cannot go on top of the plate. In future, the robot could plan an order for stacking objects. At the start of the rearrangement, the robot could spread out all the objects on the table to reduce occlusions as it detects all the objects it needs to arrange.

Robustness of cross-domain object alignment. We use pre-trained semantic features from ImageNet, inspired by [46], to align real and generated objects. However, the generated images sometimes lack detail, e.g. the generated keyboards lack legible text, making alignment difficult. As diffusion models improve, this issue will be mitigated.

B. Future Work

Personal preferences. If objects placed by users are visible in the inpainting mask, DALL-E may implicitly condition images on inferred preferences (e.g. left/right-handedness). Future work could extend to conditioning on preferences inferred in previous scenes arranged by users [8].

Prompt engineering. Adding terms such as "neat, precise, ordered, geometric" for the dining scene improved the apparent neatness of the generated image. As found in other works [47], there is significant scope to explore this further.

Language-conditioned rearrangement. User instructions can easily be added to the text prompt, e.g. "plates stacked" vs "plates laid out". Prior work shows that following spatial relations such as "inside of" is difficult for some diffusion models [48], but future work could overcome this.

C. Recommendations to the Text-To-Image Community

As this is the first work to explore web-scale diffusion models for robotics, we now provide our findings on how future diffusion models can be made more useful for robotics.

Everyday scenes in training datasets. We found that Stable Diffusion [29] trained on LAION-Aesthetics is proficient at generating aesthetically pleasing images, but the DALL-E training approach may be better suited for robotic applications, because the training dataset includes a significant amount of "ordinary" images and stock photographs. Training *only* on everyday photographs could be useful.

Visual conditioning. Rather than just conditioning on language descriptions of objects to be generated, it would be useful to condition on image features of the real objects, but still allow the diffusion model to arrange them differently. This would help with matching between the initial and generated images. Techniques such as [49], [50] can make the generated objects better match the real instances.

Activity-oriented datasets. Building web-scale models which feature activities that we would like robots to perform could lead to breakthroughs in robotics. Text-to-video models [51], [52], [53] can be used as powerful world models. Even text-to-image models trained on frames from videos involving everyday activities can be useful.

3D geometry. Extracting 3D geometry from web-scale models trained on 2D image data [54] can allow for 6-DoF object pose estimation, making robotics methods such as DALL-E-Bot applicable to 3D scenes, e.g. stacking shelves.

D. Conclusions

In this paper, we show for the first time that web-scale diffusion models like DALL-E have significant potential as "imagination engines" for robots, acting like an aesthetic prior for arranging scenes in a human-like way. This allows for zero-shot, open-set, and autonomous rearrangement, using DALL-E without requiring any further data collection or training. In other words, **our system gives web-scale diffusion models an embodiment to actualise the scenes that they imagine**. Studies with human users showed that they are happy with the results for everyday rearrangement tasks, and that the inpainting feature of diffusion models is useful for conditioning on pre-placed objects. We believe that this is an exciting direction for the future of robot learning, as diffusion models continue to impress and inspire complementary research communities.

ACKNOWLEDGEMENTS

Research presented in this paper has been supported by Dyson Technology Ltd, and the Royal Academy of Engineering under the Research Fellowship Scheme. The authors would like to thank Andrew Davison, Ignacio Alzugaray, Kamil Dreczkowski, Kirill Mazur, Eric Dexheimer, and Tristan Laidlow for helpful discussions. The authors would also like to thank Kentaro Wada for developing some of the robot control infrastructure which was used in the experiments.

REFERENCES

[1] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," *arXiv*, 2022.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.

[4] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan, and A. Zeng, "Do as I can, not as I say: Grounding language in robotic affordances," *arXiv*, 2022.

[5] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, P. Sermanet, N. Brown, T. Jackson, L. Luu, S. Levine, K. Hausman, and B. Ichter, "Inner monologue: Embodied reasoning through planning with language models," *arXiv*, 2022.

[6] D. Batra, A. X. Chang, S. Chernova, A. J. Davison, J. Deng, V. Koltun, S. Levine, J. Malik, I. Mordatch, R. Mottaghi, M. Savva, and H. Su, "Rearrangement: A challenge for embodied AI," *arXiv*, 2020.

[7] Y. Jiang, M. Lim, and A. Saxena, "Learning object arrangements in 3D scenes using human context," *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, 2012.

[8] I. Kapelyukh and E. Johns, "My house, my rules: Learning tidying preferences with graph neural networks," in *Conference on Robot Learning (CoRL)*, 2021.

[9] W. Liu, C. Paxton, T. Hermans, and D. Fox, "StructFormer: Learning spatial structure for language-guided semantic rearrangement of novel objects," *2022 International Conference on Robotics and Automation (ICRA)*, 2022.

[10] V. Jain, Y. Lin, E. Undersander, Y. Bisk, and A. Rai, "Transformers are adaptable task planners," in *6th Annual Conference on Robot Learning*, 2022.

[11] Y. Kant, A. Ramachandran, S. Yenamandra, I. Gilitschenski, D. Batra, A. Szot, and H. Agrawal, "Housekeep: Tidying virtual households using commonsense reasoning," *arXiv*, 2022.

[12] G. Sarch, Z. Fang, A. W. Harley, P. Schydlo, M. J. Tarr, S. Gupta, and K. Fragkiadaki, "TIDEE: Tidying up novel rooms using visuo-semantic commonsense priors," in *European Conference on Computer Vision*, 2022.

[13] M. J. Schuster, D. Jain, M. Tenorth, and M. Beetz, "Learning organizational principles in human environments," in *2012 IEEE International Conference on Robotics and Automation*, 2012, pp. 3867–3874.

[14] N. Abdo, C. Stachniss, L. Spinello, and W. Burgard, "Robot, organize my shelves! Tidying up objects by predicting user preferences," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[15] Y. Lin, A. S. Wang, E. Undersander, and A. Rai, "Efficient and interpretable robot manipulation with graph neural networks," *IEEE Robotics and Automation Letters*, vol. 7, pp. 2740–2747, 2022.

[16] C. Finn and S. Levine, "Deep visual foresight for planning robot motion," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[17] M. Babaeizadeh, M. T. Saffar, S. Nair, S. Levine, C. Finn, and D. Erhan, "FitVid: Overfitting in pixel-level video prediction," *arXiv*, 2020.

[18] M. Wu, F. Zhong, Y. Xia, and H. Dong, "TarGF: Learning target gradient field for object rearrangement," *arXiv*, 2022.

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

[20] M. Shridhar, L. Manuelli, and D. Fox, "CLIPort: What and where pathways for robotic manipulation," in *Proceedings of the 5th Conference on Robot Learning (CoRL)*, 2021.

[21] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani, and J. Lee, "Transporter networks: Rearranging the visual world for robotic manipulation," *Conference on Robot Learning (CoRL)*, 2020.

[22] A. Goyal, A. Mousavian, C. Paxton, Y.-W. Chao, B. Okorn, J. Deng, and D. Fox, "IFOR: Iterative flow minimization for robotic object rearrangement," *arXiv*, 2022.

[23] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

[24] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, 2020.

[25] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," 2021.

[26] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Advances in Neural Information Processing Systems*, 2021.

[27] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv*, 2021.

[28] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," *arXiv*, 2022.

[29] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *arXiv*, 2021.

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

[31] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," *ArXiv*, 2021.

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, 2021.

[33] P. Besl and H. McKay, "A method for registration of 3-D shapes. ieee trans pattern anal mach intell," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 1992.

[34] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[35] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[36] A. Gupta, P. Dollar, and R. Girshick, "LVIS: A dataset for large vocabulary instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[37] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, "OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[38] A. Akbik, D. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling," in *COLING 2018, 27th International Conference on Computational Linguistics*, 2018.

[39] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "FLAIR: An easy-to-use framework for state-of-the-art NLP," in *NAACL 2019, 2019 Annual Conference of the North*

American Chapter of the Association for Computational Linguistics (Demonstrations), 2019.

[40] W. Goodwin, S. Vaze, I. Havoutis, and I. Posner, "Semantically grounded object matching for robust robotic scene rearrangement," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022.

[41] H. W. Kuhn and B. Yaw, "The Hungarian method for the assignment problem," *Naval Res. Logist. Quart*, 1955.

[42] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "ImageNet-21K pretraining for the masses," *arXiv*, 2021.

[43] T. Ridnik, H. Lawen, A. Noy, E. Ben, B. G. Sharir, and I. Friedman, "TResNet: High performance gpu-dedicated architecture," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[44] V. Vosylius and E. Johns, "Where to start? Transferring simple skills to complex environments," in *6th Annual Conference on Robot Learning*, 2022.

[45] E. Johns, "Coarse-to-fine imitation learning: Robot manipulation from a single demonstration," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[46] K. Mazur, E. Sucar, and A. J. Davison, "Feature-realistic neural fusion for real-time, open set scene understanding," *arXiv*, 2022.

[47] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *ArXiv*, 2022.

[48] C. Conwell and T. Ullman, "Testing relational understanding in text-guided image generation," *arXiv*, 2022.

[49] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv*, 2022.

[50] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," *arXiv*, 2022.

[51] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang, "CogVideo: Large-scale pretraining for text-to-video generation via transformers," *arXiv*, 2022.

[52] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, and Y. Taigman, "Make-A-Video: Text-to-video generation without text-video data," *arXiv*, 2022.

[53] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, and T. Salimans, "Imagen video: High definition video generation with diffusion models," *arXiv*, 2022.

[54] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "DreamFusion: Text-to-3D using 2D diffusion," *arXiv*, 2022.