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

Anonymous Author(s)
Affiliation
Address
email

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. Crucially, we show this is possible zero-shot using only the pre-trained DALL-E model, without needing any further data collection or training. Encouraging real-world results with human studies show that this is an exciting direction for using these web-scale pre-trained models in robot learning algorithms. We also propose a list of recommendations to the text-to-image community, to align further development of these models with applications to robotics. Videos are available at: sites.google.com/view/dallebot

Keywords: Diffusion Models, Image Generation, Object Rearrangement

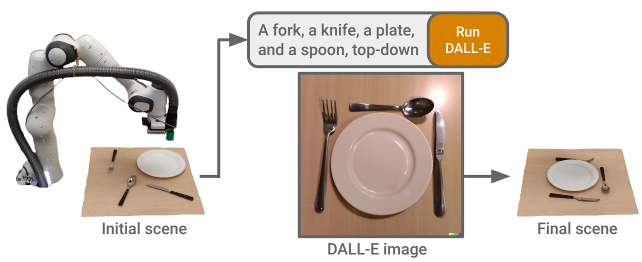

Figure 1: The robot prompts DALL-E with the list of objects it detects, which generates an image with a human-like arrangement of those objects. The robot then recreates that arrangement in reality.

1 Introduction

Diffusion models such as DALL-E [1] have recently shown an astonishing ability to generate high-quality images from text prompts, by training on hundreds of millions of captioned images from the web [2, 3, 4]. Previous breakthroughs in web-scale foundation models have been applied successfully to robotics [5, 6, 7, 8, 9]. In this work, we explore the following question: **How can image diffusion models such as DALL-E, pre-trained on web-scale data, be used for 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 consider that they could be used to generate goal images for general object rearrangement tasks [10], such as setting a table, loading a dishwasher, tidying a room, stacking a shelf, and assembling furniture. Most prior methods for predicting the goal state (i.e. a goal pose for each object) require manually collecting a dataset of examples for how a scene should be arranged [11, 12, 13, 14, 15, 16, 17, 18]. Our proposed framework predicts how to arrange a given

Submitted to the 6th Conference on Robot Learning (CoRL 2022). Do not distribute.

scene without requiring this data collection, which restricts most existing methods to a specific set of objects and scenes. Further analysis of prior work can be found in Appendix A.

In this paper, we propose DALL-E-Bot, the first method to use web-scale image diffusion models for robotics. We design a framework which takes an image of the initial, unorganised scene, uses DALL-E to imagine a human-like goal image for that scene, and creates the corresponding object arrangement with a real robot (Fig. 1). Experiments show that this can be applied to several everyday rearrangement tasks to create arrangements which are satisfactory to humans. Additionally, we find that DALL-E's inpainting feature can precisely predict the poses of missing objects in a scene, conditioned on the pre-placed objects. Furthermore, we present a discussion of the method's limitations in Appendix J, and in Appendix K we propose ideas for future web-scale diffusion models to maximise their usefulness for robotics. Videos are available at: sites.google.com/view/dallebot

Using web-scale image diffusion models for predicting rearrangement goal states has several strengths. First, this is a *zero-shot* transfer of the DALL-E model to the object rearrangement task, because it uses the publicly available DALL-E without any additional data collection or training. Second, this is an *open-set* method: it is not restricted to a specific set of objects, because of the web-scale training of DALL-E. Third, this pipeline is *autonomous*: no human effort is required from the user, because there is no need for a human-created goal image or language guidance.

2 Method

We address the problem of predicting the goal state of a rearrangement task, i.e. a goal pose for each object, such that the objects are arranged in a natural and human-like way. The method must predict this goal state from a single RGB image I_I of the initial scene. We achieve this through a modular approach shown in Fig. 2. At the heart of our method is a web-scale image diffusion model DALL-E 2 [1], which generates high-quality samples of goal images I_G with human-like object arrangements using a language description of the scene y extracted from the initial observation.

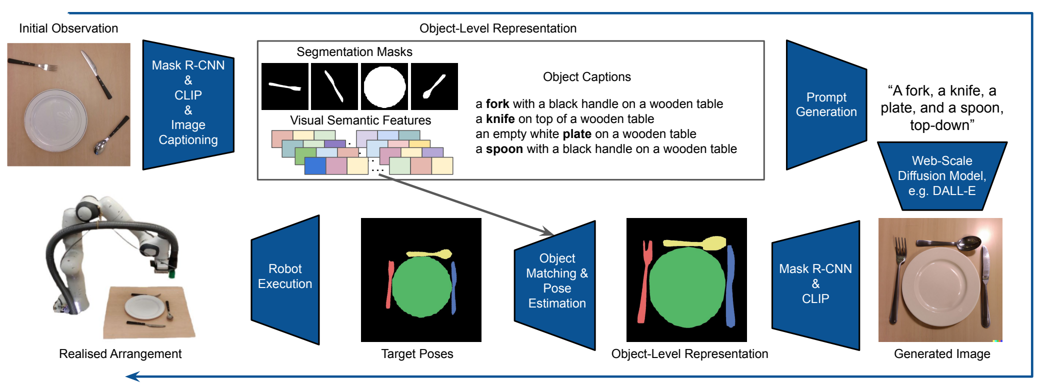

Figure 2: An overview of our method's pipeline.

First, we need to convert an initial RGB observation into a more relevant object-level representation to reason about the objects in the scene and their arrangement. We do so by constructing a representation that consists of text captions of crops of individual objects c_i in the scene together with their segmentation masks M_i and visual-semantic feature vector v_i acquired using the CLIP model [19].

We use text captions c_i to automatically construct a text prompt containing a list of the objects in the scene. We also append the term "top-down" so that DALL-E generates images from the same perspective as the initial image captured by a camera mounted on a robot's wrist pointing downwards better. In addition, we generate an image mask I_M that prevents DALL-E from altering the pixels corresponding to the contours of stationary objects (i.e. an object that the robot is not allowed to move) and tabletop edges to avoid objects being generated on the edge of the image.

We generate several images with the goal arrangement by sampling a conditional distribution $p_\theta(I_G|y, I_M)$ represented by a web-scale text-to-image diffusion model DALL-E 2 [1]. We convert generated images into object-level representations and filter out the ones that do not contain the same number of objects as the initial scene. From the remaining images, we select the one that minimises the cost of the linear sum assignment problem (Hungarian matching) between the object-level visual-semantic feature vectors in the initial and generated images.

Using Iterative Closest Point (ICP) [20], we then register corresponding segmentation masks to obtain transformations that need to be applied to the objects to achieve the goal arrangement. To account for possible size differences for the same object in initial and generated images, we move objects closer together or further apart, but do not allow them to collide. Finally, we convert these transformations from image to Cartesian space using a depth camera observation and deploy a real Franka Emika Panda robot equipped with a suction gripper to arrange the objects. More detailed explanations of each component in our method can be found in Appendices B-E.

3 Experiments

3.1 Zero-Shot Autonomous Rearrangement

In our experiments, we evaluate the ability of our method to create human-like arrangements using both subjective (Section 3.1) and objective (Section 3.2) metrics. First, we explore the following question: **can DALL-E-Bot arrange a set of objects in a human-preferred way?**

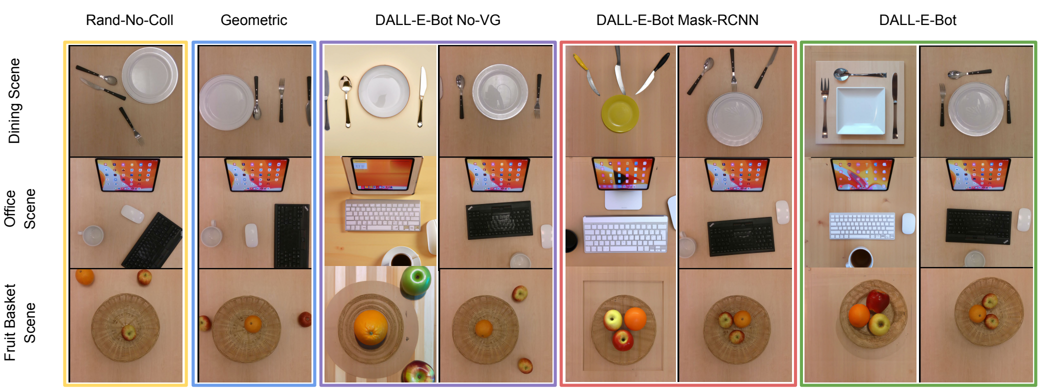

Figure 3: Examples of scenes arranged by the robot via different methods. Columns for the methods that use DALL-E include the generated image (left) and an image of the final arrangement (right).

We evaluate on 3 everyday tabletop rearrangement tasks: **dining**, **office**, and **fruit basket** (Fig. 3). The robot should arrange the objects in a human-like way while considering the poses of stationary objects (the fruit basket and the iPad), which the robot is not allowed to move. Further setup details are in Appendix F. Since DALL-E-Bot is the first method to predict precise goal states for rearrangement zero-shot, we need to design baselines which are also zero-shot for a fair comparison. We use heuristic baselines and ablation variants of DALL-E-Bot, detailed in Appendix G.

As we aim to create human-preferred arrangements, we evaluated each method by showing human participants images of the final scene created by the robot. Participants were asked: *"If the robot made this arrangement for you, how happy would you be?"*, with ratings on a Likert Scale from 1 (very unhappy) to 10 (very happy). We recruited 17 participants with ages

Method	Dining Scene	Office Scene	Fruit Scene	Mean
Rand-No-Coll	2.21±2.24	3.42±2.47	3.26±2.00	2.96
Geometric	4.14±1.80	3.47±2.23	2.85±1.62	3.49
Abl-Mask-RCNN	3.99±2.60	7.53±2.04	5.86±3.40	5.79
Abl-No-VG	**7.45**±1.81	7.13±1.99	5.48±3.56	6.69
DALL-E-Bot	7.14±2.13	**7.71**±2.01	**9.55**±0.90	**8.13**

Table 1: User ratings for the arrangements made by each method. Each figure represents the mean and standard deviation across all users and scene initialisations.

ranging from 20 to 71 (ten male, six female, and one preferred not to say). Each rated the results of 5 methods on 5 random initialisations of 3 scenes, for a total of 1275 ratings. Results are in Table 1. DALL-E-Bot beats the heuristic baselines, showing that **people value semantic correctness over simple geometric alignment**. The ablation results justify our design decisions to use object-level captioning and visual grounding. For a detailed analysis, please see Appendix I.

3.2 Placing Missing Objects with Inpainting

In the next experiment, we use objective metrics to answer the 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 user-made scene. We use the dining scene because it has the most rigid structure for semantic correctness and thus is most suitable for quantitative, objective evaluation. To create these scenes initially, we recruited ten participants (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 single object in the scene, we asked the users to provide any alternative poses of each object individually that they would still be happy with while keeping other objects fixed. We show example arrangements in Appendix H.

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

Table 2: Position and orientation errors between predicted and user-made object poses. Median is presented across all users.

We start with the image of the arrangement made by a user, and mask out everything except the 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 a 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 as the missing object, and across all the users. We use two zero-shot heuristic methods as baselines, detailed in Appendix G. For each method, 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 2. DALL-E-Bot outperforms the heuristic baselines, and is able to accurately place the missing objects with a small error across the different users. This implies that it is successfully conditioning on the placement of the other objects in the scene using inpainting, and that the human and robot can create an arrangement collaboratively.

3.3 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 how to arrange objects in a human-like way. This allows for zero-shot, autonomous rearrangement, using DALL-E out-of-the-box, without requiring collecting datasets of example arrangements for specific scenes, and without any additional training. In other words, our system gives web-scale diffusion models an embodiment to realise the scenes that they imagine. Studies with human users showed that they are happy with the created arrangements for everyday rearrangement tasks, and that the inpainting feature of diffusion models is useful for conditioning on pre-placed objects. Web-scale diffusion models are a recent and active research frontier, and so we have also provided recommendations for further aligning these models with robotics. We believe that this is an exciting direction for the future of robot learning, as pre-trained diffusion models continue to impress and inspire complementary research communities.

References

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

[2] 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, 2021.

[3] 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, 2022.

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

[5] 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.

[6] 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, 2022.

[7] Y. Cui, S. Niekum, A. Gupta, V. Kumar, and A. Rajeswaran. Can foundation models perform zero-shot task specification for robot manipulation? In *Proceedings of The 4th Annual Learning for Dynamics and Control Conference*, 2022.

[8] 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. In *arXiv preprint arXiv:2207.05608*, 2022.

[9] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell. Real-world robot learning with masked visual pre-training. *CoRL*, 2022.

[10] 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, 2020.

[11] 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.

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

[13] V. Jain, Y. Lin, E. Undersander, Y. Bisk, and A. Rai. Transformers are adaptable task planners, 2022.

[14] 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.

[15] 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.

[16] 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.

[17] M. Kang, Y. Kwon, and S.-E. Yoon. Automated task planning using object arrangement optimization. In *2018 15th International Conference on Ubiquitous Robots (UR)*, pages 334–341, 2018.

[18] A. Taniguchi, S. Isobe, L. E. Hafi, Y. Hagiwara, and T. Taniguchi. Autonomous planning based on spatial concepts to tidy up home environments with service robots. *Advanced Robotics*, 35 (8):471–489, 2021.

[19] 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*, Proceedings of Machine Learning Research, 2021.

[20] 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.

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

[22] 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*, pages 3867–3874, 2012.

[23] 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.

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

[25] M. Wu, F. Zhong, Y. Xia, and H. Dong. Targf: Learning target gradient field for object rearrangement, 2022.

[26] A. Goyal, A. Mousavian, C. Paxton, Y.-W. Chao, B. Okorn, J. Deng, and D. Fox. Ifor: Iterative flow minimization for robotic object rearrangement. In *arXiv:2202.00732*, 2022.

[27] 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.

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

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

[30] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

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

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

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

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

[35] 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.

[36] 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.

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

[38] 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.

[39] 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.

[40] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916, 2022.

[41] C. Conwell and T. Ullman. Testing relational understanding in text-guided image generation, 2022.

[42] 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, 2022.

A Related Work

A.1 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 [21], and out-of-place objects can be detected automatically [16]. At a room level, the correct drawer or shelf can be classified [22], taking preferences into account [23]. Lower-level prediction from a dense set of goal poses can be achieved with a graph neural network [24] or a preference-aware transformer [13]. Our framework generates high-resolution images of how objects should be placed, thus not requiring a set of discrete options to be pre-defined, and providing more precise guidance than with language-based methods.

Learning to predict continuous object poses can be done using example arrangements by encoding spatial preferences with a graph VAE [12], or using an autoregressive language-conditioned transformer [14], or by learning gradient fields [25]. Other methods use full demonstrations [5, 15], or leverage priors such as human pose context [11]. When the goal image is given, rearrangement is possible even with unknown objects [26]. 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. It also does not require a human user to complete the rearrangement task themselves in order to provide a goal image. Instead, exploiting existing web-scale image diffusion models enables zero-shot, autonomous rearrangement.

A.2 Web-Scale Diffusion Models

Generating images with web-scale diffusion models such as DALL-E is at the heart of our method. Diffusion models [27] 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 [28, 29, 30], text-conditioned images [1, 2, 3, 4], robot trajectories [31], and audio waves [32]. We use DALL-E 2 [1] in this work, although our framework could be used with other text-to-image models.

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 [33] from the Detectron2 library [34] to detect objects in an image and generate segmentation masks $\{M_i\}_i^n$. This model was pre-trained on the LVIS dataset [35], which has 1200 object classes, being more than sufficient for many rearrangement tasks. The Mask R-CNN model provides us with object bounding boxes, their segmentation masks and class labels. However, while bounding box and segmentation mask predictions are usually high-quality (regardless of the predicted class), and can be used for pose estimation (described in Section E), the assigned class labels are often incorrect due to the large number of classes in the training dataset.

As we are using text labels of objects in the scene (described in Section C) to construct a prompt for an image diffusion model, it is crucial for these labels to be accurate and descriptive. Instead of directly using predicted object class labels, we pass RGB crops of each object individually through the OFA image-to-text captioning model [36] and acquire a text description of the objects in the initial scene observation $\{c_i\}_i^n$. Generally, this approach allows us to more accurately predict object class labels and go beyond the objects in the training distribution and even obtain their visual characteristics such as colour, material and shape.

Finally, we also pass each object crop through a CLIP visual model [19], 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. Thus we have converted an initial scene RGB observation I_I into an object-level representation of the scene $\{M_i, c_i, v_i\}_i^n$, with a segmentation mask, a text caption, and a semantic feature vector for each object.

C Goal Image Generation

Our method relies on the ability to generate images of natural and human-friendly arrangements given their language descriptions. To this end, we heavily utilise the recent advances in text-to-image generation using web-scale diffusion models. Specifically, we use the DALL-E 2 [1] model from OpenAI. It was trained on a vast number of image-caption pairs from the Web, and represents the conditional distribution $p_\theta(I_G|y, I_M)$. Here, I_G is an image generated by the model, y 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 friendly and usable way. Therefore, by sampling this distribution, we can generate images representing our desired scenes and realise the object arrangements by matching the object poses in them. Additionally, the ability to condition this distribution on image mask I_M allows us to tackle scenarios where not all objects in the scene need to or can be moved by the robot.

We first need to construct a text prompt y describing the desired scene. To this end, we use object captions from our object-level representation. Although full captions, including their visual characteristics, could be used to generate images with identical objects in the scene, in this work, we only use the nouns describing the object's class and leave including visual characteristics for future work. 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 [37] from the Flair NLP library [38], 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. This work deals with tabletop scenes with initial observations captured by a camera mounted on a robot's wrist pointing downwards. Therefore, we added a "top-down" phrase to the prompt to better align the initial and generated images. We have also found that it reduces the frequency of generated images with unusual, artistic camera perspectives. An example prompt we use would be "A fork, a knife, a plate, and a spoon, top-down".

We use the 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). Secondly, 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 and incentivises DALL-E to create objects of appropriate sizes. Finally, we subtract enlarged segmentation masks of all the movable objects from I_M to avoid any shadows. The latter is essential, as if DALL-E sees any shadows of objects in their original poses, it will generate objects in the same poses to match the shadows, hindering the method's performance.

Using the prompt y and the conditional mask I_M, we sample a batch of images from the conditional distribution $p_\theta(I_G|y, I_M)$, represented by 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 generated images, not all will be desirable for the rearrangement task: some may have artefacts which make object detection difficult, others may contain the wrong number of objects, etc. We need to select the generated image I_G which best matches 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 same procedure as in Section B. We filter out generated images with the wrong number of objects, compared to the initial scene. Then, we match the objects in the generated image to the objects in the initial image. This is non-trivial since the generated objects are different instances to the real objects, with a very different appearance. Inspired by [39], a similarity score between any two objects (one from I_I, and one from I_G) is computed 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 to compute an assignment of each object in the live 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 contains the most similar set of objects to the real scene, and so that arrangement is most likely to transfer well to the real objects.

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 the segmentation masks, we can estimate a transformation from the initial pose (in the initial image) to the goal pose (in the generated image). We rescale the initial segmentation masks to match the corresponding ones in the generated image and use the Iterative Closest Point algorithm [20] to align the two masks, taking each pixel to be a point. This gives us a 3-DoF (x, y, θ) transform in pixel space which would move an object from its pose in the initial image to its goal pose in the generated image. The scale of the objects in the generated image can be significantly different, leading to the found arrangement resulting in collisions or being un-naturally spaced out. Therefore, we adjust the poses of the objects in the scene based on the size difference of objects in the initial and generated images. We do so by moving the objects closer or further from the object with the minimum cumulative distance to all the other objects while also pushing objects out of any undesired collisions.

Next, we use the depth camera to project the pixel-space poses into 3D space on the tabletop, obtaining a transformation for each object which moves it from the initial pose to the goal pose. Finally, we use the real robot to realise these transformations using a suction gripper. It moves the gripper to the object using inverse kinematics. The object is grasped with the suction gripper using a grasping primitive. The robot rotates the object while it is being transported to the goal pose, also using inverse kinematics and motion planning. The robot then places the object. It also reasons about the execution of the whole task and moves objects that would cause collisions into intermediate placement locations if needed before moving them to their predicted goal poses.

F Evaluation Setup

The **dining scene** involves four objects (a knife, a fork, a spoon, and a plate), and a robot should be able to arrange them so that a user would be happy seeing said arrangement when sitting down for a meal. The **office scene** includes a stationary object (a display) and three movable objects (a keyboard, a mouse and a mug). The arrangement of movable objects should be natural and useable with respect to the stationary object that a robot cannot move. Finally, the **fruit basket scene** contains two apples and an orange, as well as a stationary basket. This scene is challenging because it requires reasoning about the spatial relations between the fruits and the basket, and because the fruit in the generated images is often densely packed partially occluding the basket. The rearrangements are executed on a Franka Emika robot equipped with a compliant suction gripper. We record the outcome as an RGB image of a tabletop captured by RealSense D435i mounted on the wrist of the robot.

G Baselines

G.1 Zero-Shot Autonomous Rearrangement

Since DALL-E-Bot is the first method to predict arrangements zero-shot, we devised additional training-free methods as baselines, which can create arrangements that are natural to humans in our evaluation scenes. The Rand-No-Coll arrangement strategy arbitrarily places objects in the environment while ensuring they do not overlap. The Geometric baseline puts all the objects in a horizontal line such that they are not colliding, and the longer side of the object-oriented bounding box is aligned with the y-axis. In addition, we compare our method DALL-E-Bot against two of its ablation variants, Abl-Mask-RCNN and Abl-No-VG, to showcase the importance of accurate text description of objects in the scene and visual grounding of DALL-E predictions (described in Section C). The former uses labels predicted by Mask R-CNN to construct the text prompt for DALL-E, while the latter does not use the edge crops around the table.

G.2 Placing Missing Objects with Inpainting

Hand-designed baselines (Rand-No-Coll and Geometric) aim to place the missing object in a geometrically pleasing way based on the poses of other objects in the scene.

The Rand-No-Coll approach places the missing object arbitrarily in the workspace, ensuring it does not collide with the fixed objects. The Geometric baseline places the object on a line defined by centroids of segmentation maps of two fixed objects while also matching the alignment of the closest object.

The distribution of acceptable poses is multimodal, which can cause significant errors if a method finds a mode not selected by the user. Therefore, we present the median across all users, which is less dominated by outliers than the mean, so it is a better representation of the aggregate performance.

H User-Provided Arrangements for Inpainting

Figure 4: Example arrangements made by users for the inpainting experiment.

In the inpainting experiment, we ask users to create example arrangements so that methods can predict the poses of masked-out objects. In Fig. 4, we visualise several of the example arrangements provided by users. Even for a scene with as much semantic structure as a dining table, there is still significant variation in how users arrange this scene, due to their national cultural background or personal preferences. This shows that the methods benefit from conditioning on the placement of the pre-placed objects in order to place the missing object correctly. It also justifies our evaluation methodology for handling this multi-modal distribution, where we ask the users to provide several example placements for an object if they consider them all acceptable, and methods should predict any of these to achieve a low error.

I Experimental Results Discussion

I.1 Zero-Shot Autonomous Rearrangement

Looking at the user studies results presented in Table 1 in the main paper, we can see that DALL-E-Bot receives higher user scores, showing that it can create satisfactory arrangements even without task-specific training. It beats the heuristic baselines, showing that users do care about semantic correctness for arranging scenes beyond just geometric alignment, and justifying the use of web-scale learning for capturing these subtle semantic arrangement rules. This is especially evident in the fruit scene, where DALL-E recognises the semantic connections between fruit and a fruit basket. Since it has seen many paintings and photographs of fruit in fruit baskets, it successfully predicts that this is a natural goal state. The Abl-Mask-RCNN [33] ablation baseline falls short on the dining scene, since it often predicts the wrong classes for the objects, e.g. frisbee instead of plate. This makes the prompt to DALL-E unusual, resulting in unnatural generated arrangements. This justifies our use of a dedicated captioning model instead of Mask R-CNN classes. DALL-E-Bot also outperforms the Abl-No-VG ablation baseline on some scenes, showing that there is some value in including the edges of the scene in the inpainting mask, which makes DALL-E less likely to generate objects only partially in the image. Qualitative results in Fig. 3 in the main paper also show that the realised arrangements for DALL-E-Bot closely match the diffusion-generated images for most objects, despite the differences between object instances.

I.2 Quantitative Evaluation

As we can see from Table 2 in the main paper, considered baselines struggle with finding the correct placements of the missing object. This shows that it is challenging to design a method for this task without overfitting to one specific object. On the other hand, DALL-E-Bot can consistently infer and estimate the preferred pose of the missing object by only observing the fixed objects. Note that each component in the pipeline will contribute to the end-to-end error, e.g. due to imperfect segmentation or pose estimation. Since our method is modular, it is easy to swap in another component, e.g. a more powerful pose estimator if object models are available, and decrease the error in this way. Challenges like instance segmentation are independent and active areas of research: as new state-of-the-art models are developed, they can easily be integrated into our method to improve its performance.

J Limitations & Future Work

Here, we discuss the limitations of this method to help researchers decide whether it is well-suited for their use-case, and propose ideas for future work.

Personal preferences. If objects placed by the user are visible in the inpainting mask, DALL-E may implicitly condition images on inferred preferences (e.g. left/right-handedness). However, when no objects are pre-placed by the user, then the arrangement made by the robot may not align with the user's preferences. Future work could extend to conditioning on preferences inferred from previous scenes arranged by the user.

Top-down scenes. Our experiments focus on 3-DoF top-down scenes. This is sufficient for many common rearrangement tasks, such as setting restaurant tables or tidying office desks. Future work can extend this to 6-DoF poses, in order to e.g. stack shelves. Since the framework is modular, a 6-DoF pose estimator can easily be swapped in. It may be challenging to fit object models to the generated images.

Object-centric framework. Our method reasons about pose transformations to solve everyday rearrangement tasks. Thus, as individual components (e.g. segmentation, pose estimation) improve, overall performance will also improve. However, some tasks, such as folding deformable fabrics or sweeping small particles, are not within this method's scope.

Beyond rearrangement. This works focuses on object rearrangement, which covers many useful everyday tasks. In principle, this framework can be extended to tasks beyond object rearrangement by learning robot policies which reach the generated goal images.

Overlap between objects. Currently, our method assumes that movable objects cannot overlap, so the fork cannot go on top of the plate. To handle this, the robot would need to use task planning to stack objects in the correct order.

Robustness of cross-instance object alignment. To estimate each object's transformation from the initial image to the generated image, we align the two binary segmentation masks with ICP. However, this is difficult for symmetric objects such as knives or keyboards, which can be aligned with 180-degree error. This can be alleviated by aligning semantic feature maps instead of binary segmentation masks.

Diffusion model accessibility. We use the public-facing interface for DALL-E from OpenAI. Although this is a paid API, there are already diffusion models such as Stable Diffusion [4] which are freely available and can be used for inference in seconds on a consumer-grade GPU. As more diffusion models become widely available, it will be feasible for any research lab or company to apply these diffusion models in their robotics setup.

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 [40], there is significant scope to explore this further.

Language-conditioned generation. One exciting direction for future work is generating arrangements based on language instructions. This may prove challenging, since prior work [41] has shown that DALL-E finds it difficult to bind textual relations to objects reliably. However, this may be overcome with future diffusion models. Note also that our method does not rely on specifying relations through text, so this does not present a limitation of the current method, but an important issue to consider for future work.

K 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 the strengths and limitations of existing diffusion models for robotics, with the aim of guiding the text-to-image community when targetting applications to robotics.

Everyday scenes in training datasets. We found that Stable Diffusion [4] trained on LAION-Aesthetics is proficient at generating aesthetically pleasing images, but DALL-E is better suited for robotic applications, because the training dataset includes "ordinary" images. Taking this further, training *only* on everyday scenes could be important for robotics.

Batch sampling and rejection. Many of the generated images are not suitable as goal images (e.g. wrong number of objects). We found that the best images came from sampling larger batches and designing an algorithm to reject invalid samples. Diffusion model systems and tools which allow for automated rejection based on the text prompt 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. Textual inversion [42] looks promising for making the generated objects better match the real instances.

Outpainting. We found that objects are frequently generated at the image edge and only partially in view, making pose estimation more difficult. Outpainting, available in some text-to-image models including DALL-E, can help with this.

Guidance scale. Some interfaces (e.g. Stable Diffusion [4]) allow a trade-off between image realism and text prompt adherence. This is useful for robotics, since generating images that adhere to the text prompt is much more important than generating attractive or realistic images.