Scaling Robot Learning with Semantically Imagined Experience

Anonymous CVPR submission

Paper ID *****

Abstract

*Recent advances in robot learning have shown promise in enabling robots to perform a variety of manipulation tasks and generalize to novel scenarios. One of the key contributing factors to this progress is the scale of robot data used to train the models. To obtain large-scale datasets, prior approaches have relied on either demonstrations requiring high human involvement or engineering-heavy autonomous data collection schemes, both of which are challenging to scale. To mitigate this issue, we propose an alternative route and leverage text-to-image foundation models widely used in computer vision and natural language processing to obtain meaningful data for robot learning without requiring additional robot data. We term our method **Ro**bot Learning with **S**emantically **I**magened **E**xperience (**ROSIE**). Specifically, we make use of the state of the art text-to-image diffusion models and perform aggressive data augmentation on top of our existing robotic manipulation datasets via inpainting various unseen objects for manipulation, backgrounds, and distractors with text guidance. Through extensive real-world experiments, we show that manipulation policies trained on data augmented this way are able to solve completely unseen tasks with new objects and can behave more robustly w.r.t. novel distractors.*

1. Introduction

Though recent progress in robotic learning has shown the ability to learn a number of language-conditioned tasks [4, 26, 54, 55], the generalization properties of such policies is still far less than that of recent large-scale vision-language models [7, 46, 51]. One of the fundamental reasons for these limitations is the lack of diverse data that covers not only a large variety of motor skills, but also a variety of objects and visual domains. This becomes apparent by observing more recent trends in robot learning research – when scaled to larger, more diverse datasets, current robotic learning algorithms have demonstrated promising signs towards more robust and performant robotic systems [4, 26]. However, this promise comes with an arduous challenge: it is difficult to significantly scale up diverse, real-world data collected by robots as it requires either engineering-heavy autonomous schemes such as scripted policies [28, 35] or laborious human teleoperations [4, 24]. To put it into perspective, it took 17 months and 13 robots to collect 130k demonstrations in [4]. In [28], the authors used 7 robots and 16 months to collect 800k autonomous episodes. While some works [30, 53, 67] have proposed potential solutions to this conundrum by generating simulated data to satisfy these robot data needs, they come with their own set of challenges such as generating diverse and accurate enough simulations [26] or solving sim-to-real transfer [40, 50]. Can we find other ways to synthetically generate realistic diverse data without requiring realistic simulations or data collection on real robots?

To investigate this question we look to the field of computer vision. Traditionally, synthetic generation of additional data, whether to improve the accuracy or robustify a machine learning model, has been addressed through data augmentation techniques. These commonly include randomly perturbing the images including cropping, flipping, adding noise, augmenting colors or changing brightness. While effective in some computer vision applications, these data augmentation strategies do not suffice to provide novel robotic experiences that can result in a robot mastering a new skill or generalizing to semantically new environments [1, 34, 50]. However, recent progress in high-quality text-to-image diffusion models such as DALL-E 2 [46], Imagen [51] or StableDiffusion [48] provides a new level of data augmentation capability. Such diffusion-based image-generation methods allow us to move beyond traditional data augmentation techniques, for three reasons. First, they can meaningfully augment the semantic aspects of the robotic task through a natural language interface. Second, these methods are built on internet-scale data and thus can be used zero-shot to generate photorealistic images of many objects and backgrounds. Third, they have the capability to meaningfully change only part of the image using methods such as inpainting [70]. These capabilities allow us to generate realistic scenes by incorporating novel distractors, backgrounds, and environments while reflecting the semantics of the new task or scene – essentially distilling the vast knowledge of large generative vision models into robot experience.

In this paper, we investigate how off-the-shelf image-

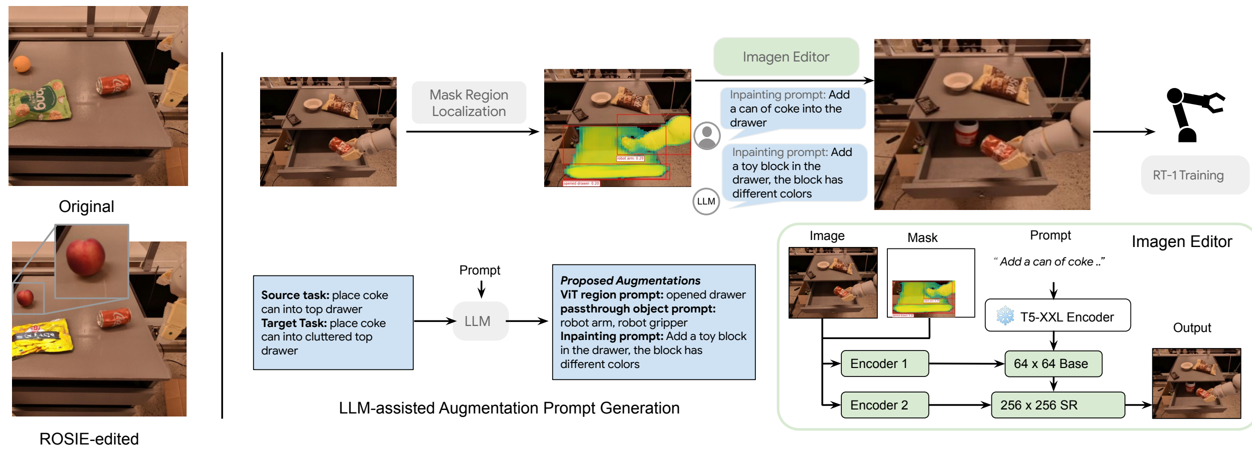


Figure 1. We propose using text-guided diffusion models for data augmentation in robot learning. These augmentations can produce highly convincing images suitable for learning downstream tasks. As demonstrated in the figure, some of the objects were produced using our system, and it is difficult to identify which are real and which are generated due to the photorealism of our system.

generation methods can vastly expand robot capabilities, enabling new tasks and robust performance. We propose **Ro**bot Learning with **S**emantically **I**magened **E**xperience **(ROSIE)**, a general and semantically-aware data augmentation strategy. ROSIE works by first parsing human provided novel instructions and identifying areas of the scene to alter. It then leverages inpainting to make the necessary alterations, while leaving the rest of the image untouched. This amounts to a *free lunch* of novel tasks, distractors, semantically meaningful backgrounds, and more, as generated by internet-scale-trained generative models. We demonstrate this approach on a large dataset of robotic data and show how a subsequently trained policy is able to perform novel, unseen tasks, and becomes more robust to distractors and backgrounds.

2. Robot Learning with Semantically Imagened Experience (ROSIE)

Our approach, ROSIE, automates robot data generation via semantic image augmentation to improve robustness and generalization of policy learning. We assume access to labeled state-action pairs of a robot performing a task with a natural language instruction. ROSIE augments the instruction with semantically different circumstances and generates masks of relevant regions. It performs inpainting with Imagen Editor based on the augmentation prompt, consistently augmenting the robot trajectory across all time steps. Details of each component are discussed in Sections 2.1 to 2.4. We use the generated data for downstream tasks such as policy learning and success detection. See Figure 1 for an overview of the pipeline.

2.1. Augmentation Region Localization using Open Vocabulary Segmentation

To generate semantically meaningful augmentations on existing robotic datasets, we detect the image region for augmentation using open-vocabulary instance segmentation. We use OWL-ViT open-vocabulary detector [41] with an additional instance segmentation head to predict fixed resolution instance masks for each bounding box detected by OWL-ViT, similar to Mask-RCNN [17]. We freeze the main OWL-ViT model and fine-tune a mask head on Open-Images-V5 instance segmentations [3, 32]. The instance segmentation model of OWL-ViT requires a language query to specify the part of the image to detect. To obtain masks for objects that the robot arm interacts with, we use the target object specified in the language instruction ℓ from each episode $\mathbf{e}$ of the robotic dataset as a prompt to perform segmentation using OWL-ViT. For example, if ℓ is "pick coke can", the target object of the task is a coke can. We also generate masks in regions where distractors can be inpainted to improve the policy's robustness. In this setting, we detect both the table and all the objects on the table using OWL-ViT. This allows us to sample a mask on the table that does not overlap with existing objects (passthrough objects). We show examples of masks detected by OWL-ViT from our robotic dataset in Figure 4.

2.2. Augmentation Text Proposal

We discuss two approaches to obtain the augmentation prompt for the text-to-image diffusion model: hand-engineered prompt and LLM-proposed prompt.

Hand-engineered prompt. The first method involves manually specifying the object to augment. To generate new tasks, we choose objects outside of our training data to expand the data support. To improve policy robustness and success detection, we randomly select semantically meaningful objects and add them to the prompt to generate meaningful distractors. For example, in Figure 3, to generate novel in-hand objects by replacing the original object (green chip bag) with various microfiber cloth, we use the prompt `Robot picking up a blue and white stripe` `cloth` to perform inpainting effectively.

LLM-proposed prompt. While hand-engineered prompt guarantees out-of-distribution data, it limits scalability. To leverage large language models (LLMs) for prompt proposal, we use GPT-3 [6] to propose objects for augmentation. We specify the original task and the target task after augmentation in the LLM prompt and ask the LLM to propose the OWL-ViT prompt for detecting masks of the target region and passthrough objects. Figure 1 shows an example of LLM-assisted augmentation prompt proposal, where LLM-generated text is informative, benefiting text-guided image editing. We use LLM-proposed prompts in our experiments, despite some noise in the prompts (see Appendix F), which generally does not affect robotic control performance.

2.3. Diffusion Model for Text-Guided Inpainting

We use Imagen Editor [66], a text-to-image diffusion model, for text-guided image editing based on a segmentation mask and an augmentation prompt. Imagen Editor is a state-of-the-art text-guided image inpainting model that is fine-tuned on a pre-trained text-to-image generator, Imagen [51], but our approach, ROSIE, is independent of the inpainting model used. Imagen Editor uses a cascaded diffusion architecture and can generate high-resolution photorealistic augmentations, which is essential for robot learning that relies on realistic images capturing physical interactions. Furthermore, Imagen Editor is trained to de-noise object-oriented masks provided by off-the-shelf object detectors [52] and random box/stroke masks [61], allowing inpainting with our mask generation procedure.

To formally summarize, given a robotic episode $\mathbf{e}=\{(\mathbf{o}_i,\mathbf{a}_i,\mathbf{o}_{i+1},\ell)\}_{i=1}^{T}$, a segmentation mask $\mathbf{m}$ indicating the target area(s) to modify, and our generated augmentation text ℓ_{aug}, we iteratively query Imagen Editor with input $\mathbf{o}_i$, $\mathbf{m}$, and ℓ_{aug} over $i=1,\ldots,T$. Imagen Editor generates the masked region according to the input text ℓ_{aug} (e.g., inserting novel objects or distractors) while ensuring consistency with the unmasked and unedited content of $\mathbf{o}_i$, resulting in the augmented image $\tilde{\mathbf{o}}_i$. If ℓ_{aug} creates a new task, we modify the instruction ℓ to $\tilde{\ell}$, as shown in Figure 3, where the original instruction $\ell=$ "pick green rice chip bag" is modified to $\tilde{\ell}=$ "pick blue microfiber cloth", "polka dot microfiber cloth," and so on. The actions $\{\mathbf{a}_i\}_{i=1}^{T}$ remain unchanged, as Imagen Editor alters novel objects consistently with the semantics of the overall image. In summary, ROSIE generates the augmented episode $\tilde{\mathbf{e}}=\{(\tilde{\mathbf{o}}_i,\mathbf{a}_i,\tilde{\mathbf{o}}_{i+1},\tilde{\ell})\}_{i=1}^{T}$. Leveraging the expressiveness of diffusion models and priors learned from internet-scale data, ROSIE provides physically realistic augmentations (e.g., Figure 2) that make robot learning more generalizable and robust, as we show in Section 3.

2.4. Manipulation Model Training

The goal of the augmentation is to improve learning of downstream tasks, e.g. robot manipulation. We train a manipulation policy based on Robotics Transformer (RT-1) architecture [4] discussed in Appendix B. Given the ROSIE augmented dataset $\tilde{\mathcal{D}}:=\{\tilde{\mathbf{e}}_j\}_{j=1}^{\tilde{N}}$, where $\tilde{N}$ is the number of augmented episodes, we train a policy on top of a pre-trained RT-1 model [4] (35M parameters, trained for 315k steps at a learning rate of 1×10^{-4}). The finetuning uses a 1:1 mixing ratio of $\mathcal{D}$ and $\tilde{\mathcal{D}}$. We follow the same training procedure described in [4] except that we use a smaller learning rate 1×10^{-6} to ensure the stability of fine-tuning.

3. Experiments

In our experimental evaluation, we focus on robot manipulation and embodied reasoning (e.g. detecting if a manipulation task is performed successfully). We design experiments to answer the following research questions: **RQ1**: Can we leverage semantic-aware augmentation to learn completely new skills only seen through diffusion models?, **RQ2**: Can we leverage semantic-aware augmentation to make our policy more robust to visual distractors?

To answer these questions, we perform empirical evaluations of ROSIE using the multi-task robotic dataset collected in [4], which consists of $\sim$130k robot demonstrations with 744 language instructions collected in laboratory offices and kitchens. These tasks include skills such as picking, placing, opening and closing drawers, moving objects near target containers, manipulating objects into or out of the drawers, and rearranging objects. For more details regarding the tasks and the data used we refer to [4]. We include the discussion of **RQ1** below and leave **RQ2** to Appendix C.

In our experiments, we aim to understand the effects of both the augmented text and the augmented images on policy learning. We thus perform two comparisons, ablating these changes: **Pre-trained RT-1 (NoAug)**: we take the RT-1 policy trained on the 744 tasks in [4]. While pre-trained RT-1 is not trained on tasks with the augmentation text and generated objects, it has been shown to enjoy promising pre-training capability and demonstrate excellent zero-shot generalization to unseen scenarios [4] and therefore, should have the ability to tackle the novel tasks to some extent; **Fine-tuned RT-1 with Instruction Augmentation (InstructionAug)**: Similar to [69], we relabel the original episodes in RT-1 dataset to new instructions generated via our augmentation text proposal 2.2 while keeping the images unchanged. We expect this method to bring the text instructions in-distribution but fail to recognize the visuals of the augmented objects.

For implementation details and hyperparameters, please see Appendix D.

3.1. RQ1: Learning new skills

To answer RQ1, we augment the RT-1 dataset via generating new objects that the robot needs to manipulate. We evaluate our method and the baselines in the following four categories with increasing level of difficulty.

Learning to move objects near and place into generated novel containers First, we test the tasks of moving training objects near unseen containers or placing such objects into the new containers. We visualize such unseen containers in Figure 8 in Appendix E. We select the tasks "move {some object} near white bowl" and "move {some object} near paper bowl" within the RT-1 dataset, which yields 254 episodes in total. We use the augmentation text proposals to replace the white bowl and the paper bowl with the following list of objects {lunch box, woven basket, ceramic pot, glass mason jar, orange paper plate}, which are visualized in Figure 8. For each augmentation, we augment the same number of episodes as the original task.

As shown in Table 1, our ROSIE fine-tuned RT-1 policy (trained on both the whole RT-1 training set of 130k episodes and the generated novel tasks) outperforms pre-trained RT-1 policy and fine-tuned RT-1 with instruction augmentations, suggesting that ROSIE is able to generate fully unseen tasks that are beneficial for control and exceeds the inherent transfer ability of RT-1.

Learning to grasp generated unknown deformable objects Third, we test the limits of ROSIE on novel tasks where the object to be manipulated is generated via ROSIE. We pick the set of tasks "pick green chip bag" from the RT-1 dataset consisting of 1309 episodes. To accurately generate the mask of the chip bag throughout the trajectory, we run our open-vocabulary segmentation to detect the chip bag and the robot gripper as the passthrough objects so that we can filter out the robot gripper to obtain the accurate mask of the chip bag when it is grasped. We further query Imagen Editor to substitute the chip bag with a fully unknown microfiber cloth with distinctive colors (black and blue), with augmentations shown in Figure 3. Table 1 again demonstrates that ROSIE outperforms pre-trained RT-1 and RT-1 with instruction augmentation by at least 150%, proving that ROSIE is able to expand the manipulation task family via diversifying the manipulation targets and boost the policy performance in the real world.

Learning to place objects into an unseen kitchen sink in a new background To stress-test our diffusion-based augmentation pipeline, we attempted to teach the robot to place an object into a sink without ever collecting data for that task in the real world. We took all the RT-1 tasks that involved placing a can into the top drawer of a counter and used ROSIE to detect the open drawer and replace it with a metal sink using Imagen Editor. We dynamically computed the mask of the open drawer at each frame of the episode, excluding the robot arm and can from the mask. The sink made the scene completely out of the training distribution, making it challenging for the pre-trained RT-1 policy. The results in the last row of Table 1 confirm this, with ROSIE achieving a 60% success rate in placing the cans in the sink, while the RT-1 policy failed to locate the cans and achieve any success. See the first row of Figure 5 for a visualization.

Overall, through these experiments, ROSIE is shown to be capable of effectively inpainting both the objects that require rich manipulation and the target object of the manipulation policy, significantly augmenting the number of tasks in robotic manipulation. These results indicate a promising path to scaling robot learning without extra effort of real data collection.

Task Family / Text Instruction	NoAug	InstructionAug	ROSIE
Move object near novel object	0.86	0.78	**0.94**
move coke can/orange near lunch box	0.8	0.6	0.9
move coke can/orange near woven basket	0.7	0.6	0.9
move coke can/orange near ceramic pot	1.0	0.9	1.0
move coke can/orange near glass mason jar	0.9	0.8	1.0
move coke can/orange near orange paper plate	0.9	1.0	0.9
Pick up novel object	0.25	0.3	**0.75**
pick blue microfiber cloth	0.1	0.4	0.8
pick black microfiber cloth	0.4	0.2	0.7
Place object into novel container	0.13	0.25	**0.44**
place coke can into orange plastic plate	0.0	0.19	0.5
place coke can into blue plastic plate	0.25	0.06	0.38
Place object into sink	0.0	-	**0.6**
place coke can into sink	0.0	-	0.8
place pepsi can into sink	0.0	-	0.4
Pick up object in new backgrounds	0.33	-	**0.71**
pick coke can on an orange table cloth	0.0	-	0.4
pick pepsi can on an orange table cloth	0.0	-	0.7
pick coke can on an blue and white table cloth	0.2	-	0.7
pick pepsi can on an blue and white table cloth	0.8	-	0.8
pick coke can near the side of a sink	0.4	-	0.5
pick pepsi can near the side of a sink	0.3	-	0.7
pick coke can in front of a sink	0.4	-	0.9
pick pepsi can in front of a sink	0.5	-	1.0
Place object into cluttered drawer	0.38	-	**0.55**
place blue chip bag into top drawer	0.5	-	0.4
place green jalapeno chip bag into top drawer	0.4	-	0.5
place green rice chip bag into top drawer	0.4	-	0.5
place brown chip bag into top drawer	0.2	-	0.8
Pick up object (with OOD distractors)	0.33	-	**0.37**
pick coke can	0.33	-	0.37

Table 1. Full Experimental Results for ROSIE. The blue shaded results correspond to RQ1 and the orange shaded results correspond to RQ2 (discussed in Appendix C). For each task family from top to the bottom, we performed evaluations with 50, 20, 16, 10, 80, 40, and 27 episodes respectively (243 episodes in total). ROSIE outperforms **NoAug** (pre-trained RT-1 policy) and **InstructionAug** (fine-tuned RT-1 policy with instruction augmentation [69]) in both categories, suggesting that ROSIE can significantly improve the generalization to novel tasks and robustness w.r.t. different distractors.

References

[1] Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

[3] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *CVPR*, 2019.

[4] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

[8] Zoey Chen, Sho Kiami, Abhishek Gupta, and Vikash Kumar. Genaug: Retargeting behaviors to unseen situations via generative augmentation. *arXiv preprint arXiv:2302.06671*, 2023.

[9] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[10] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. *arXiv preprint arXiv:1910.11215*, 2019.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[14] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021.

[15] Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 341–346, 2001.

[16] Nicklas Hansen, Rishabh Jangir, Yu Sun, Guillem Alenyà, Pieter Abbeel, Alexei A Efros, Lerrel Pinto, and Xiaolong Wang. Self-supervised policy adaptation during deployment. *arXiv preprint arXiv:2007.04309*, 2020.

[17] Kaiming He, Georgia Gkioxari, Piotr Dollr, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[18] Daniel Ho, Kanishka Rao, Zhuo Xu, Eric Jang, Mohi Khansari, and Yunfei Bai. Retinagan: An object-aware approach to sim-to-real transfer. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10920–10926. IEEE, 2021.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[21] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017.

[22] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

[23] Stephen James, Paul Wohlhart, Mrinal Kalakrishnan, Dmitry Kalashnikov, Alex Irpan, Julian Ibarz, Sergey Levine, Raia Hadsell, and Konstantinos Bousmalis. Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12627–12637, 2019.

[24] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and

Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pages 991–1002. PMLR, 2022.

[25] Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022.

[26] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts. *arXiv preprint arXiv:2210.03094*, 2022.

[27] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning*, pages 651–673. PMLR, 2018.

[28] Dmitry Kalashnikov, Jacob Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. Mt-opt: Continuous multi-task robotic reinforcement learning at scale. *arXiv preprint arXiv:2104.08212*, 2021.

[29] Ivan Kapelyukh, Vitalis Vosylius, and Edward Johns. Dall-e-bot: Introducing web-scale diffusion models to robotics. *arXiv preprint arXiv:2210.02438*, 2022.

[30] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

[31] Ilya Kostrikov, Denis Yarats, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. *arXiv preprint arXiv:2004.13649*, 2020.

[32] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv:1811.00982*, 2018.

[33] Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. *Advances in neural information processing systems*, 33:19884–19895, 2020.

[34] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, pages 5639–5650. PMLR, 2020.

[35] Alex X Lee, Coline Manon Devin, Yuxiang Zhou, Thomas Lampe, Konstantinos Bousmalis, Jost Tobias Springenberg, Arunkumar Byravan, Abbas Abdolmaleki, Nimrod Gileadi, David Khosid, et al. Beyond pick-and-place: Tackling robotic stacking of diverse shapes. In *5th Annual Conference on Robot Learning*, 2021.

[36] Bonnie Li, Vincent François-Lavet, Thang Doan, and Joelle Pineau. Domain adversarial reinforcement learning. *arXiv preprint arXiv:2102.07097*, 2021.

[37] Weiyu Liu, Tucker Hermans, Sonia Chernova, and Chris Paxton. Structdiffusion: Object-centric diffusion for semantic rearrangement of novel objects. *arXiv preprint arXiv:2211.04604*, 2022.

[38] Zhao Mandi, Homanga Bharadhwaj, Vincent Moens, Shuran Song, Aravind Rajeswaran, and Vikash Kumar. Cacti: A framework for scalable multi-task multi-scene visual imitation learning. *arXiv preprint arXiv:2212.05711*, 2022.

[39] Ajay Mandlekar, Jonathan Booher, Max Spero, Albert Tung, Anchit Gupta, Yuke Zhu, Animesh Garg, Silvio Savarese, and Li Fei-Fei. Scaling robot supervision to hundreds of hours with roboturk: Robotic manipulation dataset through human reasoning and dexterity. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1048–1055. IEEE, 2019.

[40] Bhairav Mehta, Manfred Diaz, Florian Golemo, Christopher J Pal, and Liam Paull. Active domain randomization. In *Conference on Robot Learning*, pages 1162–1176. PMLR, 2020.

[41] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection with vision transformers. *arXiv preprint arXiv:2205.06230*, 2022.

[42] Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Pooria Poorsarvi Tehrani, Ritvik Singh, Yunrong Guo, et al. Orbit: A unified simulation framework for interactive robot learning environments. *arXiv preprint arXiv:2301.04195*, 2023.

[43] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[44] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[45] Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1, 1988.

[46] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. In *arXiv:2204.06125*, 2022.

[47] Kanishka Rao, Chris Harris, Alex Irpan, Sergey Levine, Julian Ibarz, and Mohi Khansari. Rl-cyclegan: Reinforcement learning aware simulation-to-real. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11157–11166, 2020.

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjrn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 2022.

[49] Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. *Advances in Neural Information Processing Systems*, 34:12786–12797, 2021.

[50] Fereshteh Sadeghi, Alexander Toshev, Eric Jang, and Sergey Levine. Sim2real view invariant visual servoing by recurrent control. *arXiv preprint arXiv:1712.07642*, 2017.

[51] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[52] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *CoRR*, abs/1801.04381, 2018.

[53] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019.

[54] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, 2022.

[55] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. *arXiv preprint arXiv:2209.05451*, 2022.

[56] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020.

[57] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[58] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[59] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

[60] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[61] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*, 2021.

[62] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019.

[63] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.

[64] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 969–977, 2018.

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[66] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. *arXiv preprint arXiv:2212.06909*, 2022.

[67] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9068–9079, 2018.

[68] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang,

Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11097–11107, 2020.

[69] Ted Xiao, Harris Chan, Pierre Sermanet, Ayzaan Wahid, Anthony Brohan, Karol Hausman, Sergey Levine, and Jonathan Tompson. Robotic skill acquisition via instruction augmentation with vision-language models. *arXiv preprint arXiv:2211.11736*, 2022.

[70] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[71] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR, 2020.

A. Related Work

Scaling robot learning. Given the recent results on scaling data and models in other fields of AI such as language [6, 9, 11] and vision [2, 7, 13], there are multiple approaches trying to do the same in the field of robot learning. One group of methods focuses on scaling up robotic data via simulation [22, 26, 42, 53, 55, 56, 68, 71] with the hopes that the resulting policies and methods will transfer to the real world. The other direction focuses on collecting large diverse datasets in the real world by either teleoperating robots [4, 14, 24, 39] or autonomously collecting data via reinforcement learning [27, 28, 35] or scripting behaviors [10]. In this work, we present a complementary view on scaling the robot data by making use of state-of-the-art text-conditioned image generation models to enable new robot capabilities, tasks and more robust performance.

Data augmentation and domain randomization. Domain randomization [40, 63, 64] is a common technique for training machine learning models on synthetically generated data. The advantage of domain randomization is that it makes it possible to train models on a wide variety of data to improve generalization. Domain randomization usually involves changing the physical parameters or rendering parameters (lighting, texture, backgrounds) in simulation models [16, 31, 33, 36]. Others use data augmentation to transformer simulated data to be more realistic [1, 18, 47, 50] or vice-versa [23]. Contrary to these methods, we propose to directly augment data collected in the real world. We operate directly on the real-world data and leverage diffusion models to perform photorealistic image manipulation on this data.

Diffusion models for robot control. Though diffusion models [12, 19, 20, 43, 46, 51, 57, 58, 59, 60] have become common-place in computer vision, their application to robotic domains is relatively nascent. [25] uses diffusion models to generate motion plans in robot behavior synthesis. Some works have used the ability of image diffusion models to generate images and perform common sense geometric reasoning to propose goal images fed to object-conditioned policies [29, 37]. The recent concurrent works CACTI [38] and GenAug [8] are most similar to ours. CACTI proposes to use diffusion model for augmenting data collected from the real world via adding new distractors and requires manually provided masks and semantic labels. GenAug explores the usage of depth-guided diffusion models for augmenting new tasks and objects in real-world robotic data with human-specified masks and object meshes. In contrast, our work generates both novel distractors and new tasks without requiring depth. In addition, it *automatically* selects regions for inpainting with text guidance and leverages text-guided diffusion models to generate novel, realistic augmentations.

B. Preliminaries

Diffusion models and inpainting. Diffusion models are a class of generative models that have shown remarkable success in modeling complex distributions [57]. Diffusion models work through an iterative denoising process, transforming Gaussian noise into samples of the distribution guided by a mean squared error loss. Many such models also have the capability for high-quality *inpainting*, essentially filling in masked areas of an image [15, 21, 44, 70]. In addition, such approaches can be guided by language, thus generating areas consistent with both a language prompt and the image as a whole [66].

Multi-task language-conditioned robot learning. Herein we learn vision and language-conditioned robot policies via imitation learning. We denote a dataset $\mathcal{D} := \{\mathbf{e}_j\}_{j=1}^{N}$ of N episodes $\mathbf{e} = \{(\mathbf{o}_i, \mathbf{a}_i, \mathbf{o}_{i+1}, \ell)\}_{i=1}^{T}$ where $\mathbf{o}$ denotes the observation, which correspond to the image in our setting, $\mathbf{a}$ denotes the action, and ℓ denotes the language instruction of the episode, identifying the target task. We then learn a policy $\pi(\cdot|\mathbf{o}_i, \ell)$ to generate an action distribution by minimizing the negative-log liklihood of actions, i.e. *behavioral cloning* [45]. To perform large-scale vision-language robot learning, we train the RT-1 architecture [4], which utilizes FiLM-conditioned EfficientNet [62], a TokenLearner [49], and a Transformer [65] to output actions.

C. RQ2: Robustifying manipulation policies

We investigate RQ2 with two scenarios: policy robustness w.r.t. different backgrounds and new distractors.

Unseen background. We employ ROSIE to augment the background in our training data. We perform two types of augmentations: replacing the table top with a colorful table cloth and inserting a sink on the table top. We select two manipulation tasks, “pick coke can” and “pick pepsi can” from our training set, which consists of 1222 episodes in total. We run open-vocabulary segmentation to detect the table and passthrough objects, which consist of the robot arm and the target can. To generate a diverse set of table cloth during augmentation, we query GPT-3 with the following prompt:

```
inpainting prompt: pick coke can from a red and yellow
table cloth
goal: list 30 more table cloth with different vivid
colors and styles with visual details
inpainting prompt: pick coke can from
1. Navy blue and white striped table cloth
2. White and pink polka dot table cloth
3. Mint green and light blue checkered table cloth
4. Cream and gray floral table cloth
5. Hot pink and red floral table cloth
...
```

We show the some example answers from GPT-3 in blue, which are semantically meaningful. We use Imagen Editor to replace the table top except the target can with the LLM-proposed table cloth. To inpaint a sink on the table, we follow the same procedure described in the placing objects into unseen sink task in Section 3.1 except that we inpaint the sink on the table top rather than the open drawer. We fine-tune the pre-trained RT-1 policy on both the original data and the augmented episodes with generated table cloth and metal sink. As shown in Table 1, ROSIE + RT-1 signifcantly outperforms RT-1 **NoAug** in 7 out of 8 settings while performing similarly to **NoAug** in the remaining scenario, achieving an overall 115% improvement. Therefore, ROSIE is highly effectively in robustifying policy performance under varying table textures and background.

Novel distractors. To test whether ROSIE can improve policy robustness w.r.t. novel distractors and cluttered scenes, we consider the following two tasks. First, we train a policy solely from the task "pick coke can" and investigate its ability to perform this task with distractor coke cans, which have not been seen in the 615 training episodes. To this end, we employ ROSIE to add an equal number of augmented episodes with additional coke cans on the table (see Figure 6 in Appendix E for visualizations). As shown in Table 1, RT-1 + ROSIE augmentations improves the performance over RT-1 trained with "pick coke can" data only in scenarios where there are multiple coke cans on the table.

Second, we evaluate a task that places a chip bag into a drawer and investigate its ability to perform this task with distractor objects already in the drawer, also unseen during training. This scenario is challenging for RT-1, since the distractor object in the drawer will confuse the model and make it more likely to directly output termination action. We use ROSIE to add novel objects to the drawer, as shown in Figure 7 in Appendix E and follow the same training procedure as in the coke can experiment. Table 1 shows that RT-1 trained with both the original data and ROSIE generated data outperforms RT-1 with only original data. Our interpretation is that RT-1 trained from the training data never sees this situation before and it incorrectly believes that the task is already solved at the first frame, whereas ROSIE can mitigate this issue via expanding the dataset using generative models.

D. Experiment Details

D.1 Implementation Details and Hyperparameters

We take a pre-trained RT-1 policy with 35M parameters and trained for 315k steps at a learning rate of 1×10^{-4} and fine-tune the RT-1 policy with 1:1 mixing ratio of the original 130k episodes of RT-1 data and the ROSIE-generated episodes with for 85k steps with learning rate 1×10^{-6}. We follow all the other policy training hyperparameters used in [4].

To obtain the accurate segmentation mask of the target region of augmentations, we set a threshold for filtering out predicted masks with low prediction scores of both the region of the interest and passthrough objects given by OWL-ViT. In cases where we have multiple detected masks, we always select the one with highest prediction score. Specifically, for experiments where the robot is required to pick novel objects or place objects into novel containers or move objects near unseen containers (Section 3.1), we use a threshold of 0.07 to detect the in-hand objects and the containers while using a threshold of 0.05 to detect passthrough objects, which are the robot arm and robot gripper. In experiments where the robot is instructed to place the coke can or the pepsi can into the unknown sink or pick up coke can and the pepsi can with new background , we use a threshold of 0.04 to detect the table with all objects and a threshold of 0.03 to detect the passthrough objects, which are the robot arm, robot gripper and the coke can or the blue can in this case. In experiments discussed in Sections C, we use the threshold of 0.3 to detect the table or the open drawer where we want to add new distractors.

For generating LLM-assisted prompts, we perform 1-shot prompting to the LLM. For example, in the setting of generating novel distractors in the task where we place objects into the drawer (Section C), we use the following prompt to the LLM:

```
Source task:  place pepsi can on the counter
Target task:  place pepsi can on the clutter counter
ViT region prompt:  empty counter
passthrough object prompt:  robot arm, robot gripper
inpainting prompt:  add a chip bag on the counter
Source task:  place coke can into top drawer
Target task:  place coke can into cluttered top drawer
```

and LLM generates the following prompt for detecting masks and augmentations (light blue means LLM generated):

```
ViT region prompt:  empty drawer
passthrough object prompt:  robot arm, robot gripper
inpainting prompt:  add a box of crackers in the drawer
```

which is semantically meaningful for performing mask detection and Imagen Editor augmentation. We follow this recipe of prompting for all of the tasks in our experiments.

During inpainting, we take the checkpoint of Imagen Editor 64x64 base model and the 256x256 super-resolution model trained in [66] and directly run inference to produce augmentations.

During evaluation, for the tasks that perform moving objects near novel containers and grasping unseen microfiber cloth, we perform 10 policy rollouts per new container/microfiber cloth of each method. For tasks that perform placing objects into novel containers, we perform 8 policy rollouts per new container for each method. For the task where the robot is instructed to place coke can or pepsi can into the unseen kitchen sink, for each method, we perform 5 policy rollouts for coke can and pepsi can respectively. For the task where the robot is instructed to grasp the coke can and the pepsi can in new backgrounds, we evaluate each method with 10 rollouts. For the task where the robot places the object into the cluttered drawer, we perform 10 policy rollouts per object for each method. Finally, for the task that requires the robot to pick up coke can in a scene with multiple coke cans, we perform 27 policy rollouts for each approach.

D.2 Computation Complexity

We train our policy on 16 TPUs for 1 day. For obtaining segmentation masks, we perform inference of OWL-ViT on 1 TPU for 1 hour to generate 1k episodes. During augmentation, we perform inference of Imagen Editor using 4 TPUs of the 64 x 64 base model and the 256 x 256 super-resolution model respectively for 2 hours to generate 1k episodes.

E. Examples of Augmentations

We include more visualizations of augmentations generated by ROSIE in this section. In Figure 8, we show the generated episodes of ROSIE where we inpaint novel containers in the scene, which are used in the **Learning to move objects near generated novel containers** and **Learning to place objects into generated unseen containers** experiments in Section 3.1.

In Figure 6 and Figure 7, we visualize augmented episodes with new distractors, e.g. cluttered coke cans on the table and chip bags in the empty open drawer. These augmentations correspond experiments conducted in Section C.

We also visualize the attention layers in RT-1 when training on our augmented data. As seen in Fig. 9, there are attention heads focusing on our augmented objects, which indicates the augmentation seem to be effective.

Overall, note that ROSIE is able generate semantically realistic novel objects and distractors in the manipulation setting. For example, ROSIE-generated objects typically has realistic shades on the table or the drawer, which is beneficial for training manipulation policies on top of such data.

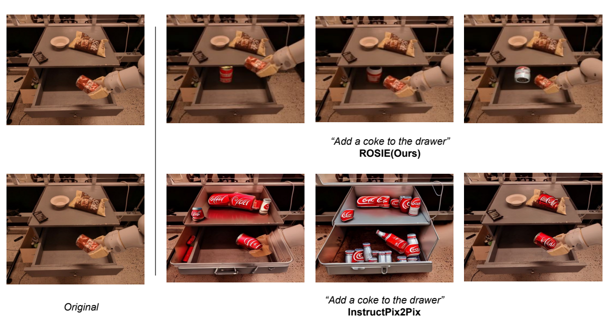


Figure 2. Our augmentation scheme generates more targeted and physically realistic augmentations that are useful for learning downstream tasks, while other text-to-image generation methods such as InstructPix2Pix [5] often makes global changes rendering the image unusable for training.

F. Failure Cases of Generated Prompts and Images

While our LLM-assisted prompts generally work very well, we would like to note that it requires few-shot prompting to work well. In the zero-shot case, LLM would just hallucinate and output unuseful augmentation prompts.

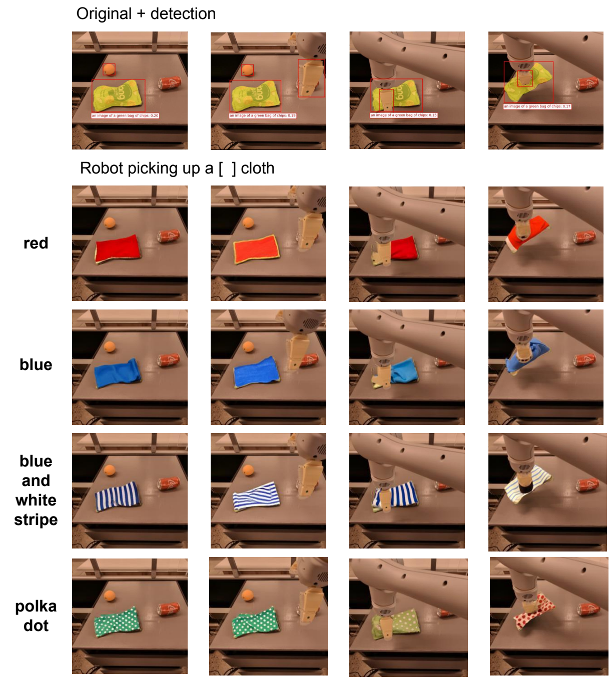


Figure 3. Augmentations of in-hand objects during manipulation. We show examples where ROSIE effectively inpaint novel objects into the original in-hand objects during manipulation. On the top row, we show the original episode with detected masks where the robot picks up the green chip bag. On the following row, we show that ROSIE can inpaint various microfiber cloth with different colors and styles into the original green chip bag. For example, we can simply pass the original episode with the masks and the prompt `Robot picking up a polka dot cloth` to get an episode the robot picking such cloth in a photorealistic manner.

For example, if we provide the following zero-shot prompt:

```
Source task:  pick coke can on a table
Target task:  pick coke can near a sink
Goal:  replace the scene in the source task with the
scene in the target task
inpainting prompt:
```

and LLM gives the following response:

```
Pick up the coke can near the sink,
replacing the one originally on the table
```

,which is not correct. Therefore few-shot prompting is crucial in ROSIE.

We show the failure cases of the augmented images in Figure 10. For the two examples on the left, ROSIE is supposed to generate woven basket and glass mason jar respectively, but it fails to generate such containers and instead generate some bowl-shape containers. For the two examples on the right, ROSIE is supposed to replace the in-hand green chip bag with blue microfiber cloth and a yellow rubber duck respectively. However, as the mask of the in-hand object becomes irregular, the performance of ROSIE degrades and ROSIE is unable to generate blue microfiber

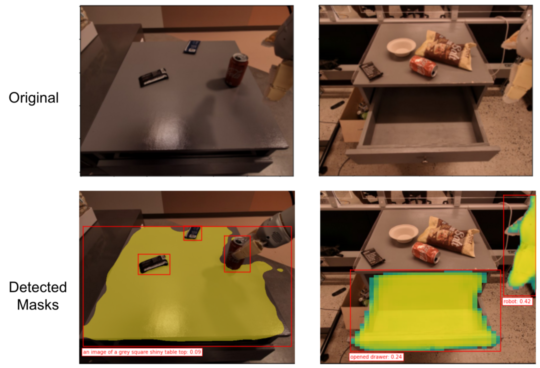


Figure 4. We show the original images from RT-1 datasets on the top row and the images with detected masks and mask labels on the bottom row.

cloth and the yellow rubber duck in full shape and half of the in-hand object remains as the green chip bag. We suspect that with fine-tuning Imagen Editor on robotic datasets that show more manipulation-related data, we can improve the generation results drastically. Note that while the generation could be suboptimal at times, our insight is that such imperfect generation can only lead to misalignment between the task instruction and images, which may not have a big negative impact on the policy results and could give extra data augmentation benefit for free. Our policy performance in Section 3 validates this insight to some degree.

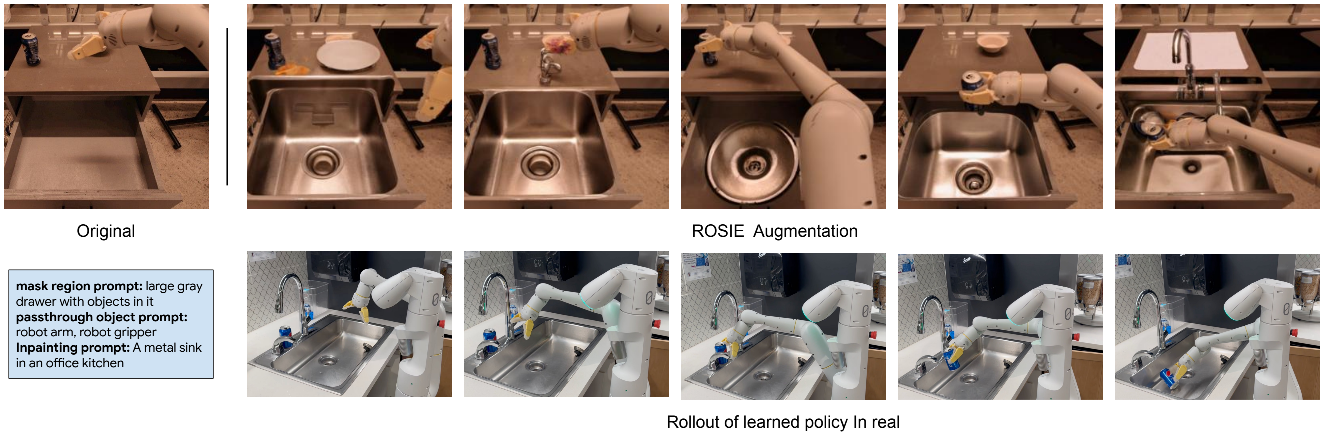
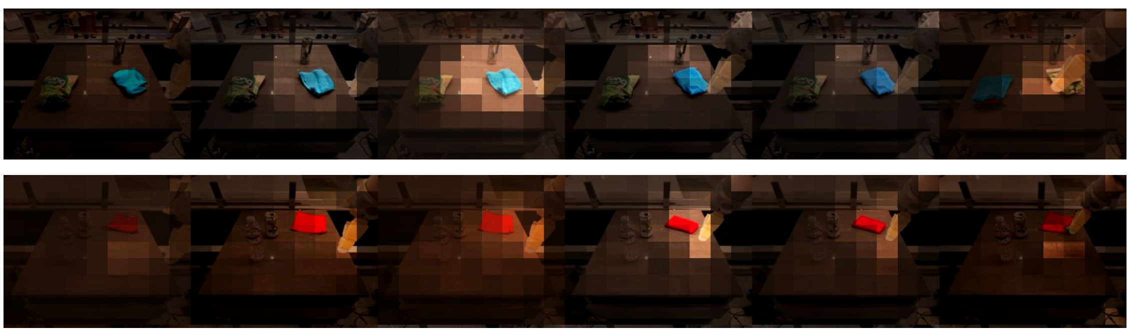


Figure 5. We show an episode augmented by ROSIE (top row) where ROSIE inpaints the metal sink onto the top drawer of the counter and a rollout of policy trained with both the original episodes and the augmented episodes in a real kitchen with a metal sink. The policy successfully performs the task "place pepsi can into sink" even if it is not trained on real data with sink before, suggesting that leveraging the prior of the diffusion models trained with internet-scale data is able to improve generalization of robotic learning in the real world.

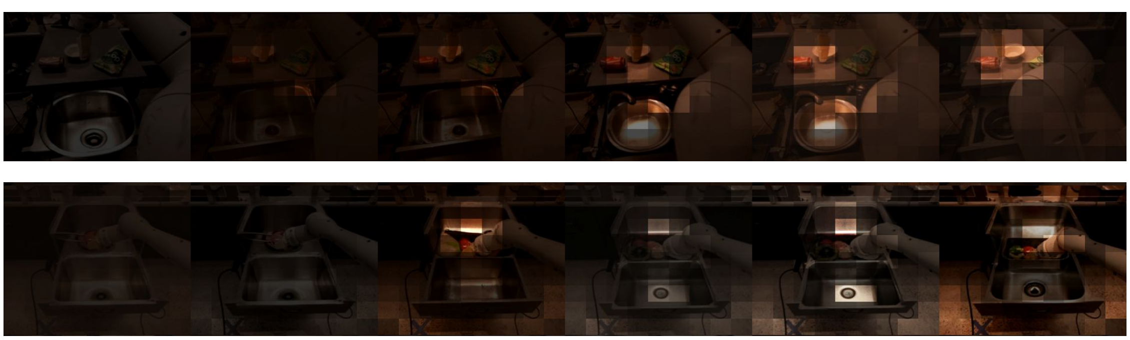
Figure 6. Augmentation Example - adding a distractor can on the table.

Figure 7. Augmentation Example - adding distractor objects into the drawer.

Figure 8. Augmentation Example - changing the container.



pick blue/red microfiber cloth



place coke in sink

Figure 9. Visualization of some attention heads focusing on our augmented objects. This visualization is an overlay of observation and the spatial attention (bright regions means high attention).

Figure 10. Failure cases of image augmentations.