Appendix

Imitation Learning and RT-1

MOO builds upon a language-conditioned imitation learning setup. The goal of language-conditioned imitation learning is to learn a policy $\pi(a \mid \ell, o)$, where a is a robot action that should be applied given the current observation o and task instruction ℓ. To learn a language-conditioned policy π, we build on top of RT-1 [24], a recent robotics transformer-based model that achieves high levels of performance across a wide variety of manipulation tasks. RT-1 uses behavioral cloning [53], which optimizes π by minimizing the negative log-likelihood of an action a given the image observations seen so far in the trajectory and the language instruction, using a demonstration dataset containing N demonstrations:

$$J(\pi) := \sum_{n=1}^{N} \sum_{t=1}^{T^{(n)}} \log \pi(a_t^{(n)} \mid \ell^{(n)}, \{o_j^{(n)}\}_{j=1}^{t}). \tag{1}$$

Vision-Language Models

In recent years, there has been a growing interest in developing models that can detect objects in images based on natural language queries. These models, known as vision-language models (VLMs), are enabling detectors to identify a wide range of objects based on natural language queries. Typically the text queries are tokenized and embedded in a high-dimensional space by a pre-trained language encoder, and the image is processed by a separate network to extract image features into the same embedding space as the text features. The language and image representations are then combined to make predictions of the bounding boxes and segmentation masks. Given a natural language query, q, and an image observation on which to run detection, o, these models aim to produce a set of embeddings for the image $f_i(o)$ with shape $(\text{height}, \text{width}, \text{feature dim})$ and an embedding of the language query $f_l(q)$ with shape feature dim such that $\text{logits} = f_i(o) \cdot f_l(q)$ gives a logit score map and is maximized at regions in o which correspond to the queries in q. Each image embedding location within $f_i(o)$ is also associated with a predicted bounding box or mask indicating the spatial extent of that object corresponding to $f_i(o)$. In this work, we use the Owl-ViT detector [54], which we discuss further in Sec. 3.4.

Datasets

We collect a focused collection of teleoperated demonstration data that focuses on increasing object diversity for the most efficient skill to collect data for, the picking task. Detailed dataset statistics across objects are shown in Appendix Figure 9.

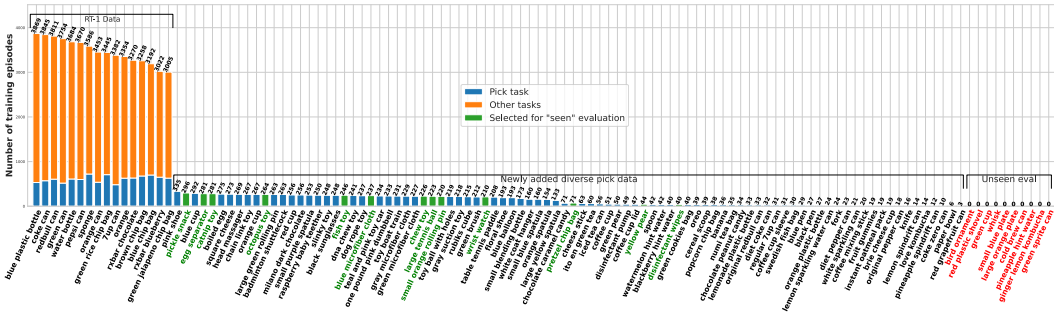

Figure 9: Distribution of training objects for "pick" episodes and other skills. The data on the left was what was used by [24]. We augmented RT-1 data with a large number of diverse pick episodes in order to demonstrate strong generalization to unseen objects. Blue and green bars represent "pick" episodes and orange bars represent other tasks like "move near" or "knock." "Green" bars were the objects we randomly selected for "seen" evaluations. All randomly selected "unseen" objects are shown to the right.

Experiments

We show a visualization of our 7-DoF manipulation robot in Figure 10.

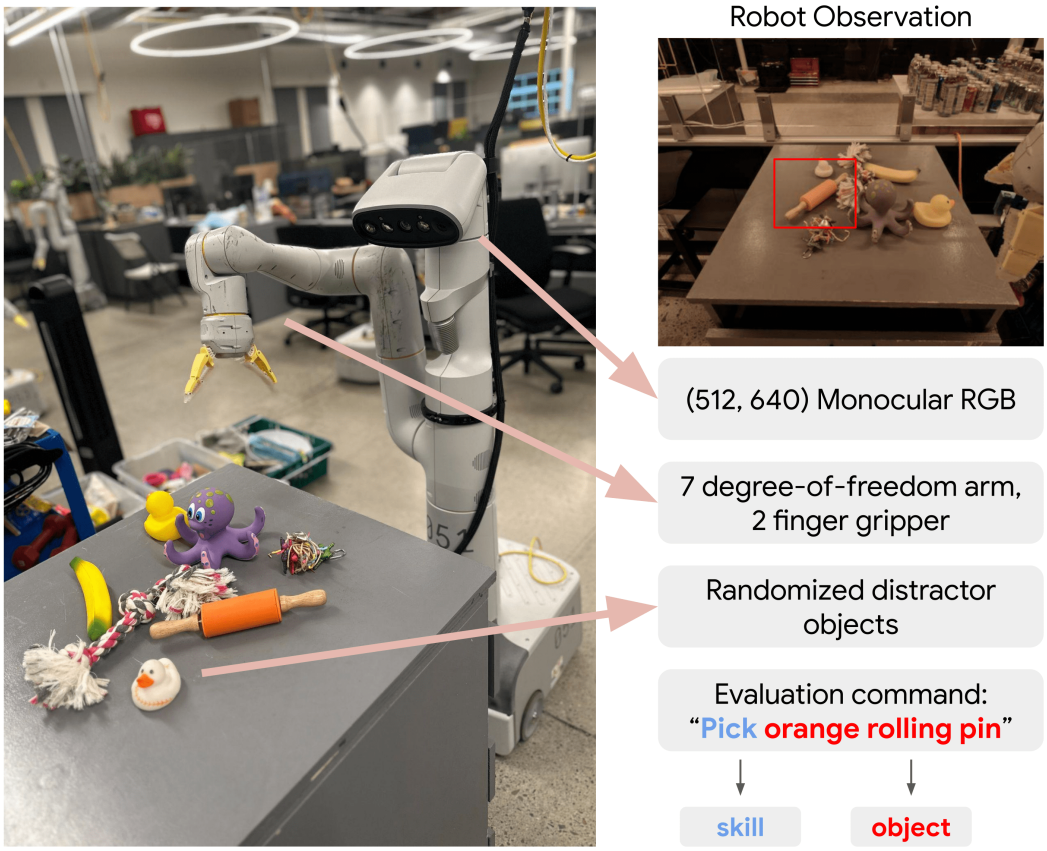

Figure 10: Image of our robot hardware and evaluation setting.

Skills. Our experiments evaluate the percent of successfully completed manipulation commands which include five skills: “pick”, “move near”, “knock,” “place upright,” and “place into” across a set of evaluation episodes. The definition of the tasks follows RT-1 [24]: For “pick”, success is defined as (1) grasping the specified object and (2) lifting the object at least 6 inches from the table top. For “move near”, success is defined as (1) grasping the specified object and (2) placing it within 6 inches of the specified target object. For “knock”, success is defined as placing the specified object from an “upright” position onto its side. “Place upright” tasks are the inverse of “knock” and involve placing an object from its side into an upright position. Finally, “place into” tasks involve placing one object into another, such as an apple into a bowl.

Robustness evaluation details. We evaluate the robustness of MOO on a variety of visually challenging scenarios with drastically different furniture and backgrounds, as shown in Figure 6; the results are reported in Figure 5. The first set of these difficult evaluation scenes introduces six evaluations across five additional open-world objects that correspond to various household items that have not been seen at any point during training. The second set of difficult scenes introduces 14 evaluations across two patterned tablecloths; these tablecloth textures are significantly more challenging than the plain gray counter-tops seen in the training demonstration dataset. Finally, the last set of difficult scenes include 14 evaluations across three new environments in natural kitchen and office spaces that were never present training. These new scenes simultaneously change the counter-top materials, backgrounds, lighting conditions, and distractor items.

Input modality demonstration details. We explore the ability of MOO to incorporate object-centric mask representations that are generated via different processes than the one used during training. During training, an OWL-ViT generates mask visual representations from textual prompts, as described in Section 3.2. We study whether MOO can successfully accomplish manipulation tasks given (1) a mask generated from a text caption from a generative VLM, (2) a mask generated from an image query instead of a text query, or (3) a mask directly provided by a human via a

Dataset Filtering		Pick	
Objects	Episodes per Object	Seen objects	Unseen objects
100%	100%	**98**	**79**
50%	100%	92	75
18%	100%	88	19
100%	50%	46	38
100%	10%	23	0

Table 1: Performance of MOO in percentage of success relative to the amount of data used for training. Both data scale and data diversity are important.

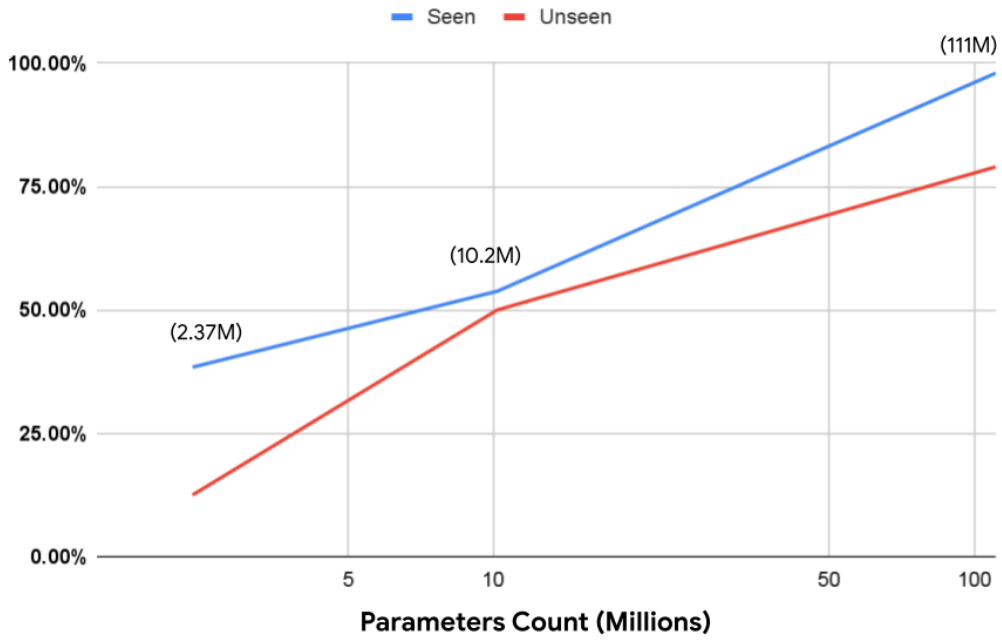

Figure 11: Pick success vs. model size. We see continuous improvements on both seen and unseen objects as we increase the number of parameters of our model architecture while keeping the data set size fixed. In comparison to our main model, we scaled down layer widths and depth by the same constant multiplier. We expect more performance gains at larger model capacity, yet are currently unable to scale further due to real time inference constraints on our robot.

GUI. For each of these cases, we implement different procedures for generating the object mask representation, which are then fed to the frozen MOO policy.

Training data ablation. We ablate the amount of data used to train MOO, and find that both data diversity and data scale are important, as shown in Table 1.

Prompts used

We use the following prompts to OWL-ViT detect our objects. All prompts were prefixed with the phrase "An image of a".
7up can → "white can of soda"
banana → "banana"
black pen → "black pen"
blue chip bag → "blue bag of chips"
blue pen → "blue pen"
brown chip bag → "brown bag of chips"
cereal scoop → "cereal scoop"
chocolate peanut candy → "bag of candy snack"
coffee cup → "coffee cup"
coke can → "red can of soda"

Figure 12: Example images of our policy detecting and grasping objects not seen during training time. The object detections are colored in correspondence to the text above the image, and the images are ordered left to right across time.

coke zero can → "can of soda"
disinfectant pump → "bottle"
fork → "fork"
green can → "green aluminum can"
green cookies bag → "green snack food bag"
green jalapeno chip bag → "green bag of chips"
green sprite can → "green soda can"
knife → "knife"
orange can → "orange aluminum can"
orange plastic bottle → "orange bottle"
oreo → "cookie snack food bag"
pepsi can → "blue soda can"
popcorn chip bag → "bag of chips"
pretzel chip bag → "bag of chips"
red grapefruit can → "red aluminum can"
redbull can → "skinny silver can of soda"
rxbar blueberry → "small blue rectangular snack food bar"
spoon → "spoon"
swedish fish bag → "bag of candy snack food"
water bottle → "clear plastic waterbottle with white cap"
white sparkling can → "aluminum can"

blue plastic bottle → “clear plastic waterbottle with white cap”
diet pepper can → “can of soda”
disinfectant wipes → “yellow and blue pack”
green rice chip bag → “green bag of chips”
orange → “round orange fruit”
paper bowl → “round bowl”
rxbar chocolate → “small black rectangular snack food bar”
sponge → “scrub sponge”
blackberry hint water → “clear plastic bottle with white cap”
pineapple hint water → “clear plastic bottle with white cap”
watermelon hint water → “clear plastic bottle with white cap”
regular 7up can → “can of soda”
lemonade plastic bottle → “clear plastic bottle with white cap”
diet coke can → “silver can of soda”
yellow pear → “yellow pear”
green pear → “green pear”
instant oatmeal pack → “flat brown pack of instant oatmeal”
coffee mixing stick → “small thin flat wooden popsicle stick”
coffee cup lid → “round disposable coffee cup lid”
coffee cup sleeve → “brown disposable coffee cup sleeve”
numi tea bag → “small flat packet of tea”
fruit gummies → “small blue bag of snacks”
chocolate caramel candy → “small navy bag of candy”
original redbull can → “can of energy drink with dark blue label”
cold brew can → “blue and black can”
ginger lemon kombucha → “yellow and tan aluminum can with brown writing”
large orange plate → “circular orange plate”
small blue plate → “circular blue plate”
love kombucha → “white and orange can of soda”
original pepper can → “dark red can of soda”
ito en green tea → “light green can of soda”
iced tea can → “black can of soda”
cheese stick → “yellow cheese stick in wrapper”
brie cheese cup → “small white cheese cup with wrapper”
pineapple spindrift can → “white and cyan can of soda”
lemon spindrift can → “white and brown can of soda”
lemon sparkling water can → “yellow can of soda”
milano dark chocolate → “white pack of snacks”
square cheese → “small orange rectangle packet ”
boiled egg → “small white egg in a plastic wrapper”
pickle snack → “small black and green snack bag”
red cup → “plastic red cup”
blue cup → “plastic blue cup”
orange cup → “plastic orange cup”
green cup → “plastic green cup”
head massager → “metal head massager with many wires”
chew toy → “blue and yellow toy with orange polka dots”
wrist watch → “wrist watch”
small orange rolling pin → “small orange rolling pin with wooden handles”
large green rolling pin → “large green rolling pin with wooden handles”
rubiks cube → “rubiks cube”
blue microfiber cloth → “blue cloth”
gray microfiber cloth → “gray cloth”
green microfiber cloth → “green cloth”
small blending bottle → “small turqoise and brown bottle”
large tennis ball → “large tennis ball”
table tennis paddle → “table tennis paddle”
octopus toy → “purple toy octopus”
pink shoe → “pink shoe”
floral shoe → “red and blue shoe”
whisk → “whisk”
orange spatula → “orange spatula”
small blue spatula → “small blue spatula”
large yellow spatula → “large yellow spatula”
egg separator → “large pink cooking spoon”

green brush → “green brush”
small purple spatula → “small purple spatula”
badminton shuttlecock → “shuttlecock”
black sunglasses → “black sunglasses”
toy ball with holes → “toy ball with holes”
red plastic shovel → “red plastic shovel”
bird ornament → “colorful ornament with blue and yellow confetti”
blue balloon → “blue balloon animal”
catnip toy → “small dark blue plastic cross toy”
raspberry baby teether → “red and green baby pacifier”
slinky toy → “gray metallic cylinder slinky”
dna chew toy → “big orange spring”
gray suction toy → “gray suction toy”
teal and pink toy car → “teal and pink toy car”
two pound purple dumbbell → “purple dumbbell”
one pound pink dumbbell → “pink dumbbell”
three pound brown dumbbell → “brown dumbbell”
dog rope toy → “white pink and gray rope with knot”
fish toy → “fish”
chain link toy → “skinny green rectangular toy”
toy boat train → “plastic toy boat”
white coat hanger → “white coat hanger”