Appendix A. Dataset Details

We give additional information about the generation of expert demonstrations in AI2-THOR, language directives, the annotation interface used to collect directives, and samples of annotations with their associated demonstrations.

A.1. Expert Demonstrations

When sampling task parameters, we employ an active strategy to maximize data heterogeneity. Figure F1 shows the distribution of high-level task across train, validation seen, and validation unseen folds. Figure F2 shows the distribution of subgoals across task types. And Figures F6 and F7 give the distributions of pickup objects and receptacles across the dataset. Each task parameter sample is defined by (t, s, o, r, m), where

- t = the task type;
- s = the scene in AI2-THOR;
- o = the object class to be picked up;
- r = the final destination for o or $\emptyset$ for **Examine**;
- m = the secondary object class for **Stack & Place** tasks ($\emptyset$ for other task types).

To construct the next tuple, we first find the largest source of imbalance in the current set of tuples. For example if $o = apple$ is more common than $o = plunger$, $o = plunger$ will be ranked higher than $o = apple$. We additionally account for the prior distribution of each entity

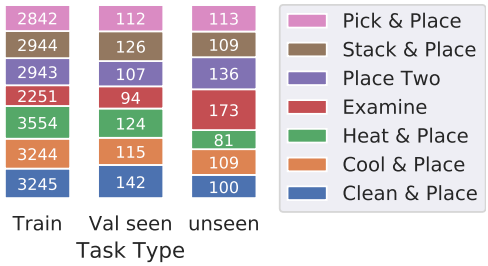


Figure F1: Task distribution across train, validation seen and unseen dataset splits.

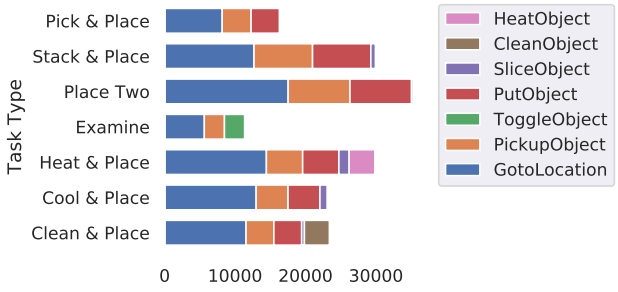


Figure F2: Subgoal distribution across 7 task types.

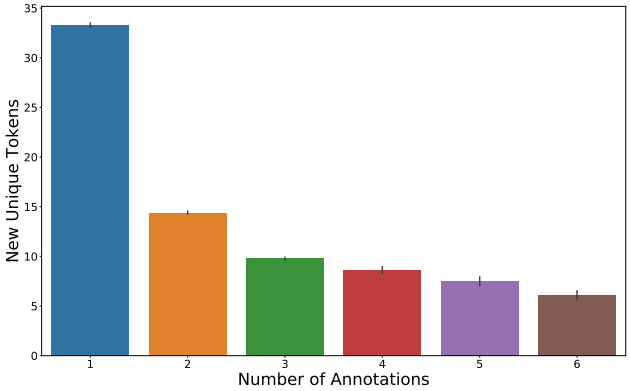


Figure F3: The number of unique tokens introduced per annotation of language directives.

(*e.g.*, if *cup* is already represented in the data often as both o and m, it becomes disfavored by the sampling algorithm for all slots). We do this greedily across all slots until the tuple is complete. Given any partial piece of information about the task, the distributions of the remaining task parameters remain heterogeneous under this sampling, weakening baseline priors such as ignoring the language input and always executing a common task in the environment.

Once a task parameter sample is complete, the chosen scene is instantiated, objects and agent start position are randomized, and the relevant room data is encoded into PDDL rules for an expert demonstration. If the PDDL planner cannot generate an expert demonstration given the room configuration, or if the agent fails an action during execution, for example by running into walls or opening doors onto itself due to physical constraints, the episode is abandoned. We gather three distinct expert demonstrations per task parameter sample. These demonstrations are further vetted by rolling them forward using our wrapper to the AI2-THOR API to ensure that a "perfect" model can reproduce the demonstration. The full sampling generation and verification code will be published along with the dataset.

A.2. Example Language Directives

We chose to gather three directives per demonstration empirically. For a subset of over 700 demonstrations, we gathered up to 6 language directives from different annotators. We find that after three annotations, fewer than 10 unique tokens on average are introduced by additional annotators (Figure F3).

A.3. Annotation Interface

Figure F4 shows the Mechanical Turk interface used to gather language annotations. Workers were presented with a video of the expert demonstration with timeline segments indicating sub-goals. The workers annotated each segment

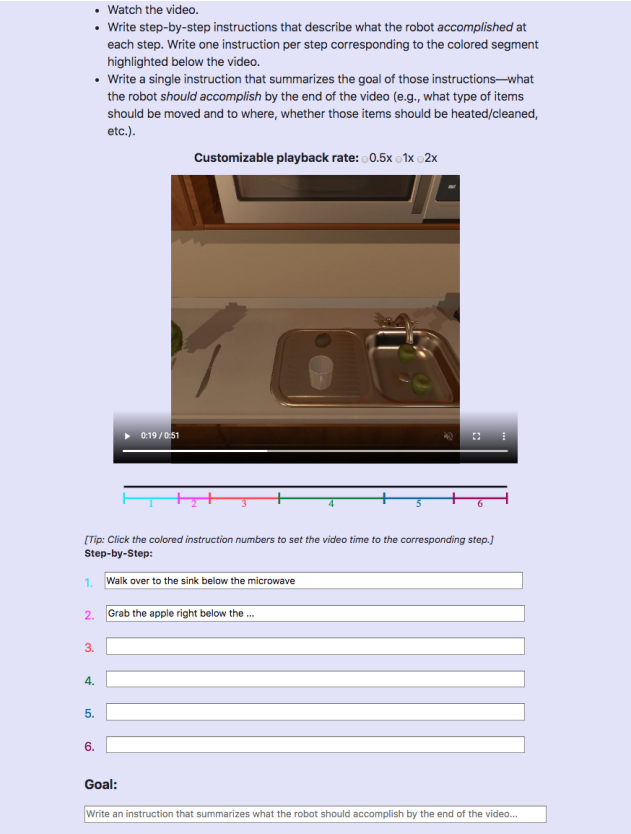


Figure F4: Mechanical Turk Annotation Interface.

while scrubbing through the video, and wrote a short summary description for the entire sequence. We payed workers \$0.7 per annotation. During vetting, annotators were paid \$0.35 per HIT (Human Interaction Task) to compare 5 sets of three directives each. These wages were set based on local minimum-wage rates and average completion time.

A.4. Vocabulary Distributions

Figure F8 shows vocabulary statistics of the language in ALFRED.

A.5. Dataset Examples

Figure F9 shows 7 expert trajectories (one per task type) and their accompanied annotations.

Appendix B. Implementation Details

We describe implementation and training details of our baseline Sequence-to-Sequence models.

Preprocessing We tokenize the language directives and convert all tokens to lower-case. During dataset generation, we save images from AI2-THOR 300×300 pixels, and later resize them to 224×224 during training. The generation

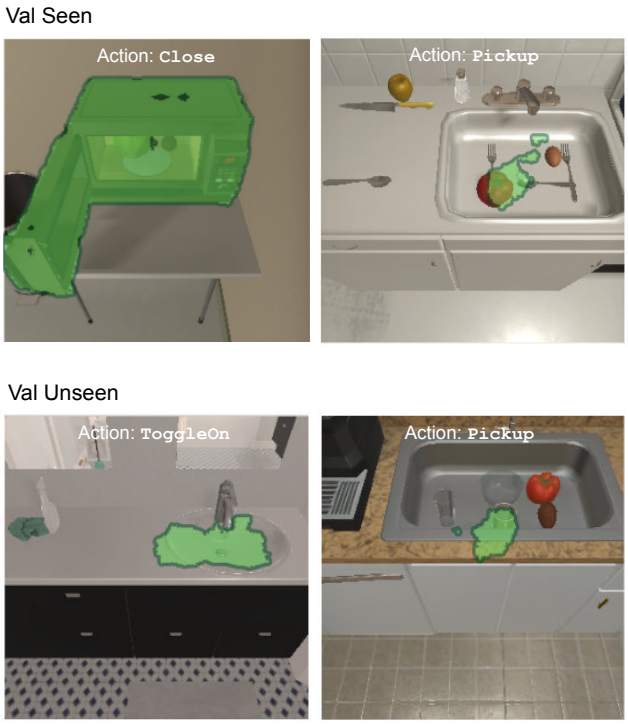


Figure F5: **Predicted interaction masks.** Masks generated by the SEQ2SEQ+PM model are displayed in green.

pipeline saves initialization information for objects and the agent, so all demonstration can be perfectly replayed in the simulator. Researchers can use this replay feature to augment the dataset by saving high-res images, depth maps, or object-segmentation masks.

Network Architecture We use a pretrained ResNet-18 [2] to extract $512 \times 7 \times 7$ features from the `conv5` layer. These features are fed into a two-layer CNN with 1×1 convolutions to reduce the channel dimension from 512 to 64. The $64 \times 7 \times 7$ output is flattened, and a fully-connected layer produces a 2500-dimensional visual feature v_t.

The language encoder is a bi-directional LSTM with a hidden-dimension of 100. We do not use pretrained language models to initialize the LSTM, and the encodings are learned from scratch in an end-to-end manner. We also use a self-attention mechanism to attend over the encodings to initialize the hidden-state of the decoder LSTM.

The action decoder is an LSTM with a hidden-dimension of 512. The actor is a fully-connected layer that outputs logits for 13 actions. The mask decoder is a three-layer deconvolution network, which takes in the concatenated vector u_t and transforms it into $64 \times 7 \times 7$ features with a fully-connected layer. These features are subsequently up-scaled into a $1 \times 300 \times 300$ binary mask through three layers of deconvolutions and up-sampling with bi-linear interpolation.

Training The models were implemented with PyTorch and trained with the Adam optimizer [3] at a learning rate of 1e-4. We use dropout of 0.3 on the visual features and the decoder hidden state, tuned on the validation data. Both the action and mask losses are weighted equally, while the auxiliary losses are scaled with a factor of 0.1. For evaluation, we choose models with the lowest loss on the validation *seen* set. It should be noted that, due to the nature of the tasks, low validation loss might not directly lead to better evaluation performance since the agent does not have to exactly imitate the expert to complete the task.

Notes on Random Agent Unlike discretized navigation where taking random actions might allow the agent to stumble upon the goal, ALFRED tasks are much harder to achieve by chance. The action space branching factor of Room-to-Room navigation [1], for example, is $4^6 \approx 4000$ (6 average steps and 4 navigation actions). By contrast, the ALFRED average branching factor is $12^{50} \approx 10^{53}$ (50 average steps for 12 actions). Beyond action type prediction, the ALFRED state space resulting from dynamic environments and the need to produce pixel-wise masks for interactive actions explodes further.

B.1. Predicted Masks

Figure F5 shows a few examples of masks generated by the SEQ2SEQ+PM model in seen and unseen validation scenes. The *Microwave* mask accurately captures the contours of the object since the model is familiar with receptacles in seen environments. In contrast, the *Sink* mask in the unseen scene poorly fits the unfamiliar object topology.

Appendix C. Additional Results

C.1. Performance by Task Type

In Table A1, we present success rates across the 7 task types. Even the best performing model, SEQ2SEQ+PM, mostly succeeds in solving some short-horizon tasks like **Pick & Place** and **Examine**. Long horizon tasks like **Stack & Place** and **Pick Two & Place** have near zero success rates across all models.

Task Ablations - **Validation**

	NO LANGUAGE		SEQ2SEQ		SEQ2SEQ+PM	
Task Type	*Seen*	*Unseen*	*Seen*	*Unseen*	*Seen*	*Unseen*
Pick & Place	0.0	0.0	6.3	**1.0**	**7.0**	0.0
Stack & Place	0.0	0.0	0.0	0.0	**0.9**	0.0
Pick Two	0.0	0.0	**1.6**	0.0	0.8	0.0
Clean & Place	0.0	0.0	0.0	0.0	1.8	0.0
Heat & Place	0.0	0.0	**1.9**	0.0	**1.9**	0.0
Cool & Place	0.0	0.0	2.4	0.0	**4.0**	0.0
Examine	0.0	0.0	4.3	0.0	**9.6**	0.0

Table A1: **Success percentages across 7 task types.** The highest values are shown in **blue**.

References

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 2018. 3

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[4] 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. 8

Figure F6: Pickup distributions in the train, validation *seen* and *unseen* folds.

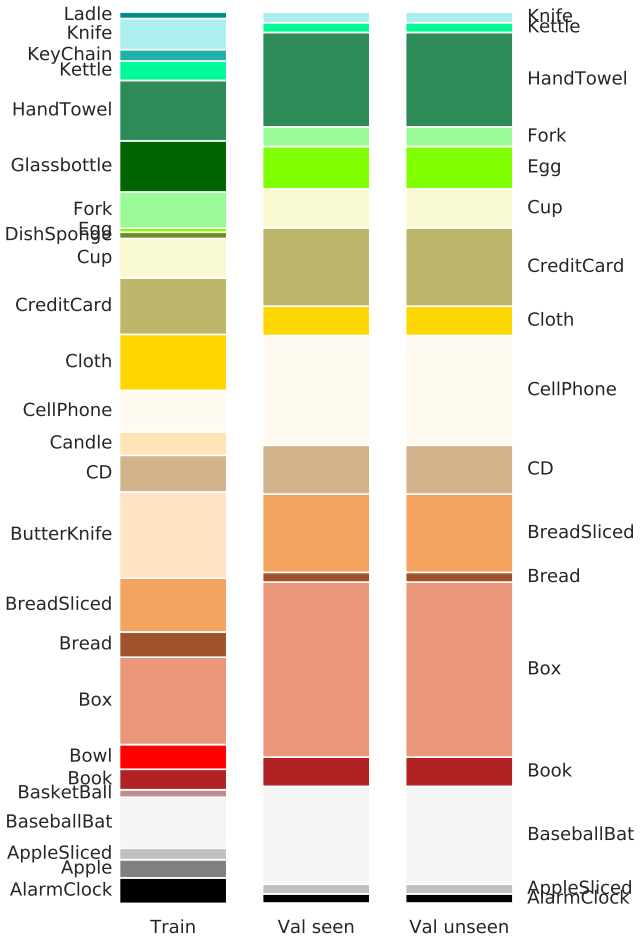


Figure F7: Receptacle distributions in the train, validation *seen* and *unseen* folds.

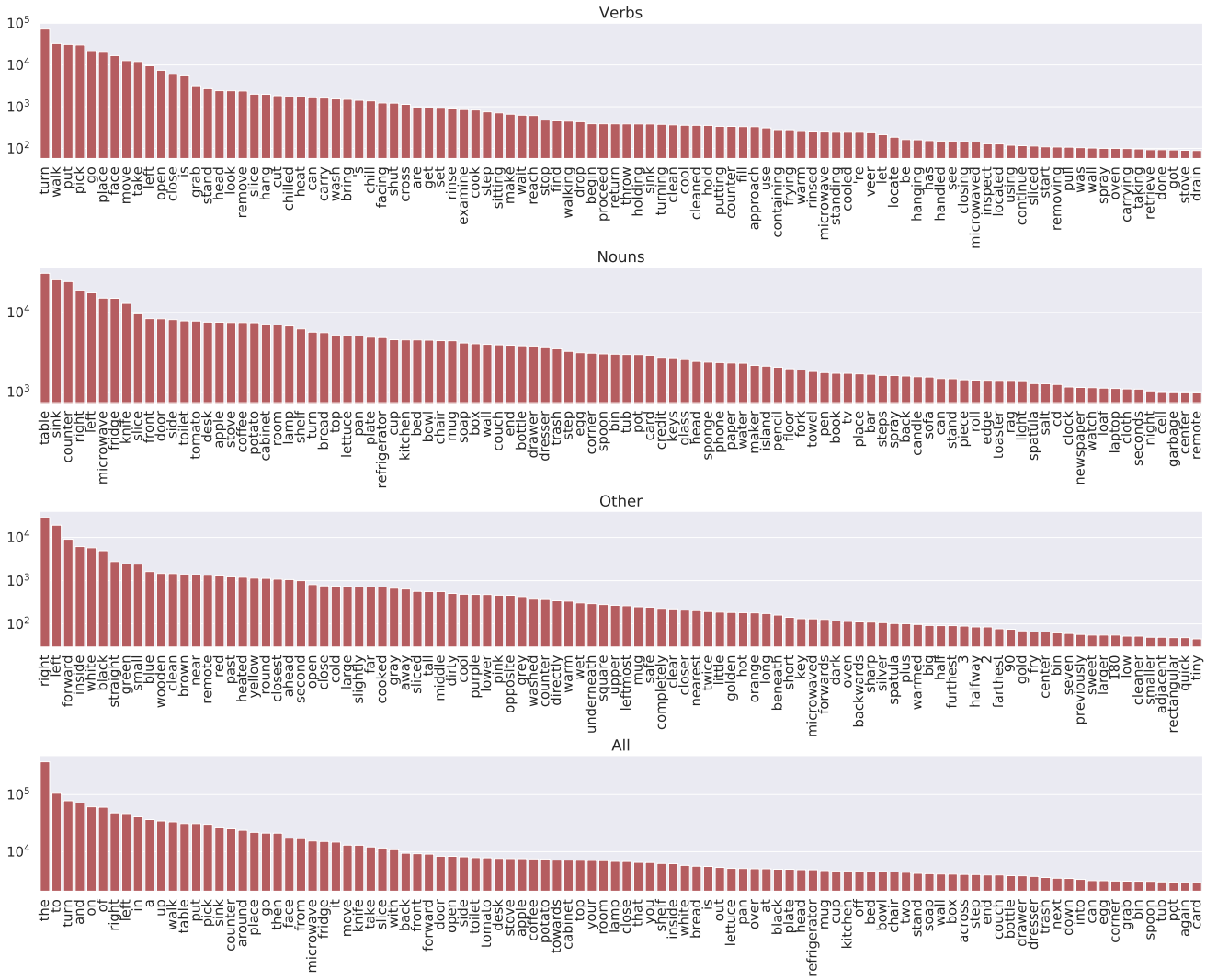
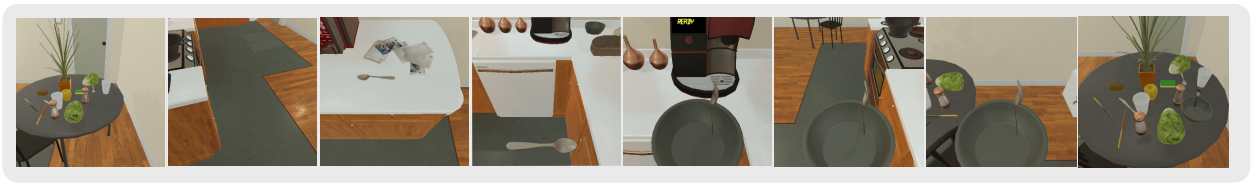


Figure F8: **Vocabulary Distributions.** Frequency distributions of 100 most common verbs, nouns, other words (non-verbs and non-nouns), and all words.

Pick & Place

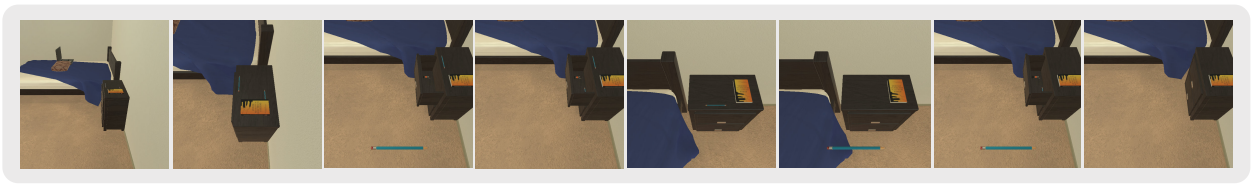
	Annotation # 1	Annotation # 2	Annotation # 3
Goals	**Put a watch on a table.**	**Put the watch on the coffee table.**	**Move the watch to the coffee table.**
Instructions	Go to the right and turn around to face the end of the cabinet with the television on it. Pick the watch up from the cabinet. Go to the right and then turn to face the coffee table in front of the couch. Put the watch on the table.	Turn right and go to the TV stand. Pick up the watch from the stand. Turn around and face the coffee table. Put the watch on the coffee table.	Turn right, go straight, turn right, step forward, then turn right to face the table with the TV on it. Pick up the watch on the table, to the right of the remote. Turn right, move forward, turn left, move forward, then turn right to face the coffee table. Put the watch on the front left corner of the coffee table.

Stack & Place



	Annotation # 1	Annotation # 2	Annotation # 3
Goals	**Put a bowl with a spoon in it on the table.**	**Move a spoon and bowl to a table.**	**To put a spoon in a bowl plus moving them to the kitchen table.**
Instructions	Turn left, and walk across the room to the microwave. Pick up the spoon. Turn left, and walk to the coffee machine .Put the spoon in the bowl. Pick up the bowl. Turn left, walk to the table, and turn left. Put the bowl on the table.	Move a spoon and bowl to a table. Go to the counter right of the microwave. Pick up the spoon from the counter. Go to the coffeemaker. Place the spoon in a bowl next to the coffeemaker. Pick up the bowl. Go to the black table. Place the bowl down on the table.	Turn left and walk across the room to face a spoon on the right end of the counter. Pick up the spoon on the end of the counter. Turn left and walk across the room to face the coffee maker on the counter. Put the spoon in the bowl to the right of the coffee maker on the counter. Pick up the bowl with a spoon in it on the counter. Turn left and walk across the room and turn left to face the kitchen table. Place the bowl with the spoon in it on the kitchen table.

Pick Two & Place



	Annotation # 1	Annotation # 2	Annotation # 3
Goals	**Putting pencils inside of a cabinet.**	**Put two pencils in a drawer.**	**Place the two pencils in the stand.**
Instructions	Walk to the bedside table in front of you. Grab the pencil that's on the table. Move slightly to the left and open the top cabinet of the bedside table. Place the pencil inside the cabinet. Face the front of the bedside table. Grab the pencil off of the bedside table. Open the top cabinet on the table. Place the pencil inside the cabinet and then close it.	Walk straight ahead to the end table. Pick up the blue pencil on the right. Walk around to the front of the end table. Put the pencil in the top drawer of the end table. Look up at the top of the end table. Pick up the pencil on the table. Look down at the drawer. Put the pencil in the drawer on top of the other pencil.	Walk to the stand next to the bed. Grab the pencil from the stand. Open the shelf inside of the stand. Place the pencil in the top shelf of the stand. Close the shelf, walk back to the stand. Grab the other pencil from the stand. Walk back to the stand next to the bed. Place the pencil in the top shelf of the stand.

Clean & Place

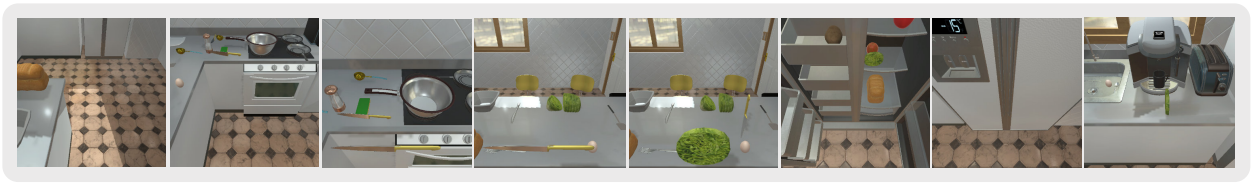
	Annotation # 1	Annotation # 2	Annotation # 3
Goals	**Put a clean rag on the top shelf of a barred rack.**	**Wash the pink towel on the shelf, put it back on the shelf.**	**Clean a red cloth.**
Instructions	Turn around, go to the barred rack. Pick up the rag from the bottle shelf of the barred rack. Go to the sink on the left. Put the rag in the sink, turn on then turn off the water. Go to the barred rack to the right of the sink. Put the rag on the top shelf of the barred rack.	Wash the pink towel on the shelf, put it back on the shelf. Turn around and go the shelf. Pick up the pink towel on the shelf. Turn around and put the towel in the sink. Fill the sink with water and wash the towel, take the towel out. Go back to the shelf. Put the towel back on the shelf.	walk on over to the towel drying rack. pick up a dirty red cloth from the towel rack. walk over to the left side of the bathroom sink. turn on the water to rinse the dirty red cloth and pick it back up again. walk back over to the towel drying rack. place the clean cloth on the drying rack.

Figure F9: **Dataset Examples.** Annotations for seven expert demonstrations.

Heat & Place

	Annotation # 1	Annotation # 2	Annotation # 3
Goals	**Put a cooked potato slice on the counter.**	**Place a slice of cooked potato onto the counter.**	**Put a piece of cooked potato on the counter.**
	Turn right, turn right, walk past the sink, turn left to face round table with tablecloth. Pick up the yellow-handled knife from the table. Cut a slice in the potato on the table. Turn left, turn left, turn right at counter, cross room, turn left at refrigerator to face counter. Put knife down on the table. Turn left, walk past sink, turn left to face round table. Pick the potato slice up from the table. Turn left, make right around corner of counter, turn left to face stove and microwave. Put potato in microwave, cook it, take it out of microwave. Turn right, cross room, turn left at counter with blue plate on it. Put potato on the counter in front of the blue plate.	Turn right, move to the table. Pick up the knife from the table. Slice the potato on the table. Turn left, move to the counter left of the bread. Put the knife on the counter near the soap container. Turn left, move to the table. Pick up a slice of potato from the table. Turn left, move to the counter in front of the stove. Put the potato slice into the microwave, cook it, pick it back up. Turn right, move to the counter left of the bread. Put the cooked potato slice on the counter.	Turn right and cross the room, then turn left and go to face the gray table. Pick up the knife from in between the lettuce and the apple. Use the knife to slice the potato that's on the gray table. Bring the knife with you and go face the kitchen counter with the loaf of bread. Put the knife down in front of the soap dispenser on the counter. Go back over to the gray table. Pick up a slice of the cut potato from the table. Bring the potato with you and go over to the stove, then look up at the microwave. Cook the potato slice in the microwave, then take it out again. Bring the potato slice over to the counter top with the loaf of bread and the knife you used to cut it. Put the potato slice down in front of the blue plate.

Cool & Place



	Annotation # 1	Annotation # 2	Annotation # 3
Goals	**Put a slice of cold lettuce on a counter.**	**Put a chilled slice of lettuce on the counter.**	**Slice some lettuce and cool it in the refrigerator so you can put it on the counter top.**
Instructions	Turn left, go forward past the counter, turn left, go forward to the counter to the left of the oven. Take the knife to the left of the large spoon from the counter. Turn around, go forward a step, turn right to the counter. Cut the lettuce on the counter into slices. Turn right, go forward a step, turn left to the counter. Put the knife behind the egg on the counter. Turn left, go forward a step, turn right to the counter. Take a slice of lettuce from the counter. Turn left, go forward, turn right to the fridge. Go to the fridge. Chill the lettuce in the fridge in front of the apple. Take the lettuce from the fridge. Turn left, go forward, turn right to face the coffee maker. Put the lettuce in front of the coffee maker on the counter.	Turn left, head toward the fridge, turn left and go to the stove. Pick up the knife beside the spoon on the counter. Turn around and turn to the right to face the counter with the egg. Cut the lettuce on the counter. Move right and back left to the counter. Put the knife behind the egg on the counter. Move the left and turn back right towards the counter. Pick up a slice of lettuce. Turn around and head to the fridge. Put the lettuce on the second shelf of the fridge, close the fridge, open the fridge and pick it back up. Turn left, go halfway across the room and turn right toward the coffee maker. Put the slice of lettuce on the counter in front of the coffee maker.	turn left and go around the counter top, then go straight to the stove top. pick up the knife with the yellow handle from behind the salt shaker on the counter top. turn left and face the counter top to your left. slice the lettuce on the counter top. face the counter top with the knife in hand. place the knife down next to the lettuce slices on the counter top. face the lettuce on the counter top. pick up a slice of lettuce on the counter top. turn left, then face the opposite wall behind you to face the refrigerator. open the refrigerator, and place the slice of lettuce in front of the apple one shelf above the bread, then shut the door and open it up again after several seconds to pick the lettuce slice up. turn left, then face forward to the part of the counter top on which the coffee maker sits. place the slice of lettuce in front of the coffee maker.

Examine in Light

	Annotation # 1	Annotation # 2	Annotation # 3
Goals	**Read a book by lamp light.**	**Examine a book with a lamp.**	**Pick up a book and turn on a lamp.**
Instructions	Head forward to the bed in front of you. Pick up the blue book that is sitting on the bed; the book that says Probabilistic Robotics. Turn to your right and walk to the night stand. Turn on the lamp that is sitting on the night stand.	walk forward a few steps, turn right, take two steps, turn left, walk to bed. pick up the book that is on the bed. turn around, take a step, turn left to face small table. turn the lamp on.	Walk forward to face the bed. Pick the book up from the bed. Turn to the right and face the night stand with the lamp. Turn the lamp on.

Figure F9: **Dataset Examples.** Annotations for seven expert demonstrations.

Figure F10: **Visual diversity of AI2-THOR [4] scenes.** Top to bottom rows: kitchens, living rooms, bedrooms, and bathrooms. Object locations are randomized based on placeable surface areas and class constraints. See https://ai2thor.allenai.org/ithor/demo/ for an interactive demo.