A. Appendix

In this appendix we provide additional results and analysis, including the state-of-the-art comparison with all evaluation metrics, the complete breakdown of performance improvements, the impact of synthetic data size, and visualizations of transformer attention maps.

A.1. Comparison with state of the art

We present a complete table comparing E.T. to the state-of-the-art methods from the public leaderboard. In addition to the success rates on validation and test folds reported in the main paper (denoted as *Full task*), we measure the amount of subgoal conditions completed for each task on average [56] (denoted as *Goal Cond.*). We also compute path-length-weighted scores [2] for both metrics which weight the original metric value by the ratio of the agent path length and the expert path length [56]. Table A1 shows that the results on the additional metrics strongly correlate with the full task success rates reported in the main paper.

Model	Validation			
	Seen		*Unseen*	
	Full task	Goal Cond.	Full task	Goal Cond.
Shridhar *et al.* [56]	3.70 (2.10)	10.00 (7.00)	0.00 (0.00)	6.90 (5.10)
Nguyen *et al.* [64]	N/A	N/A	N/A	N/A
Singh *et al.* [58]	19.15 (13.60)	28.50 (22.30)	3.78 (2.00)	13.40 (8.30)
E.T.	33.78 (24.90)	42.48 (33.10)	3.17 (1.34)	13.12 (7.41)
E.T. (pretr.)	37.63 (28.03)	47.59 (37.27)	3.76 (2.20)	14.65 (8.44)
E.T. (pretr. + & joint tr.)	**46.59 (32.26)**	**52.82 (42.24)**	**7.32 (3.34)**	**20.87 (11.31)**
Model	Test			
	Seen		*Unseen*	
	Full task	Goal Cond.	Full task	Goal Cond.
Shridhar *et al.* [56]	3.98 (2.02)	9.42 (6.27)	0.39 (0.80)	7.03 (4.26)
Nguyen *et al.* [64]	12.39 (8.20)	20.68 (18.79)	4.45 (2.24)	12.34 (9.44)
Singh *et al.* [58]	22.05 (15.10)	28.29 (22.05)	5.30 (2.72)	14.28 (9.99)
E.T.	28.77 (19.77)	36.47 (28.00)	5.04 (1.94)	15.01 (8.73)
E.T. (pretr.)	33.46 (23.82)	41.08 (31.52)	5.56 (2.82)	15.44 (9.62)
E.T. (pretr. & joint tr.)	**38.42 (27.78)**	**45.44 (34.93)**	**8.57 (4.10)**	**18.56 (11.46)**
Human	-	-	91.00 (85.80)	94.50 (87.60)

Table A1: Comparison with the models submitted to the public leaderboard on validation and test folds. We report success rates of the models on full tasks and subgoal conditions. We weight agent path lengths with expert path lengths and report path-length-weighted scores in parenthesis. The highest value per fold is shown in **blue**. 'N/A' denotes that the scores are not reported on the leaderboard or in an associated publication.

A.2. Complete performance analysis

We present a complete breakdown of performance improvements with respect to the components added to the LSTM-based baseline model proposed by Shridhar *et al.* [56]. First, we replace ImageNet visual features with features pretrained to detect objects in ALFRED as explained in Section 4.1. Next, we replace explicit pixel mask predictions with a pretrained MaskRCNN model proposed by Singh *et al.* [58]. These two components combined bring a significant improvement over the original baseline performance [56]. We then replace the LSTM model with the E.T. architecture, pretrain the language encoder of the agent to translate human language to synthetic representations, and jointly train the agent using additional 45K demonstrations to achieve the state-of-the-art performance reported in Table A1.

Components	Seen	Unseen
LSTM baseline (Shridhar *et al.* [56])	4.8	0.2
+ ALFRED detection pretraining	8.5	0.4
+ Pretrained MaskRCNN [58]	23.2	2.4
- LSTM; + Transformer (E.T.)	33.8	3.2
+ Synthetic language pretraining	37.6	3.8
+ Joint training with 45K demonstrations	**46.6**	**7.3**

Table A2: Complete breakdown of performance improvements. We report the performance of the model proposed by Shridhar *et al.* [56] and sequentially add components that improve its success rate one by one. The components include (1) visual features pretrained to detect objects in ALFRED, (2) a pretrained MaskRCNN to predict pixel masks, (3) the E.T. model, (4) language encoder pretraining on human to synthetic translation, (5) joint training with additional data.

A.3. Impact of synthetic demonstration size

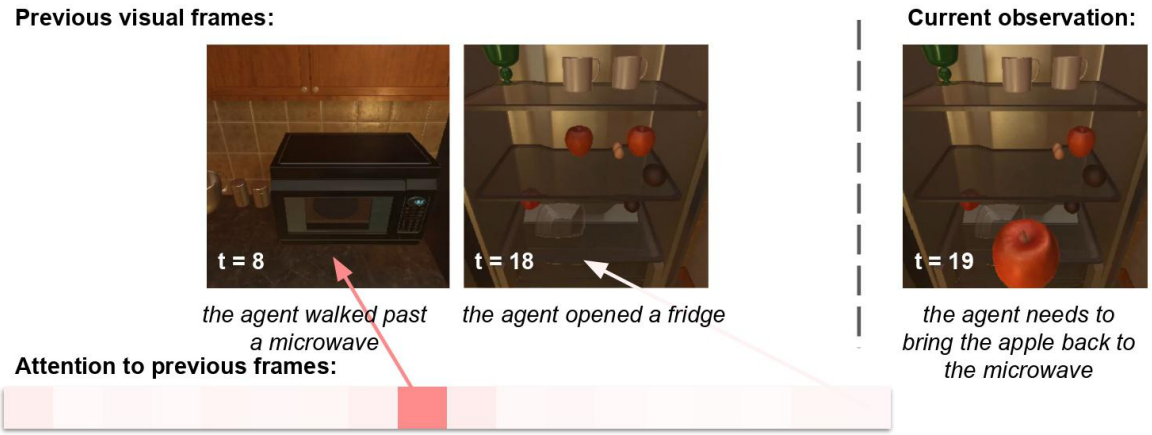
We extend the results of Table 5 and train the E.T. agent using different number of demonstrations annotated with synthetic instructions. The results are shown in Table A3. We can see that increasing the number of synthetic demonstrations in the joint training up to 22K brings a significant improvement over the model trained on human annotations only. However, doubling the synthetic demonstrations up to 44K has a very minor impact on the agent performance. We use 44K synthetic data in the main paper.

Train data	Seen	Unseen
21K human only	33.8	3.2
21K human & 11K synth.	35.5	4.1
21K human & 22K synth.	38.3	**5.5**
21K human & 44K synth.	**38.5**	5.4

Table A3: **Joint training** of the E.T. model using different number of demonstrations annotated with synthetic instructions. We report success rates on the validation folds.

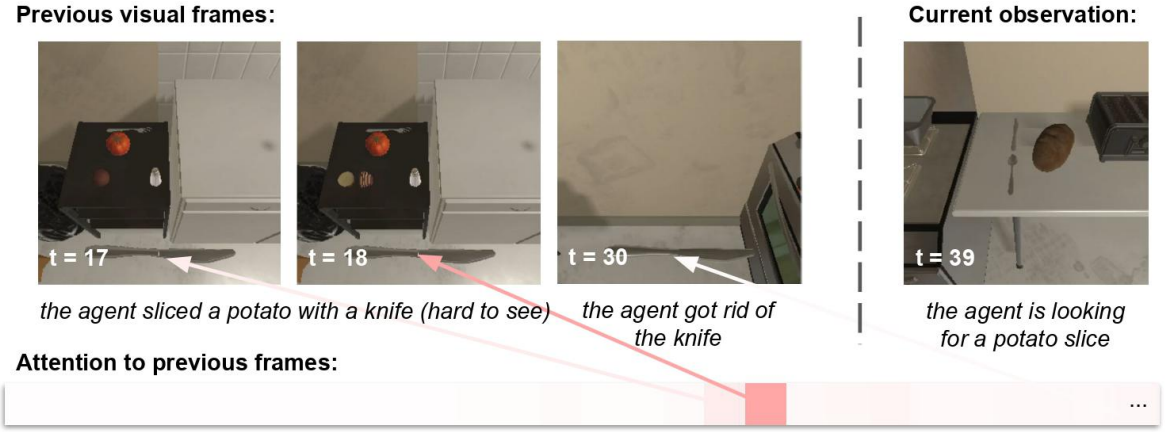
A.4. Visualizing visual attention

To better understand the impact of using a transformer encoder for action predictions, we show several qualitative examples of attention weights produced by the multimodal encoder of an E.T. agent. We use attention rollout [1] to compute attention weights from an output action to previous visual observations. Attention rollout averages attention



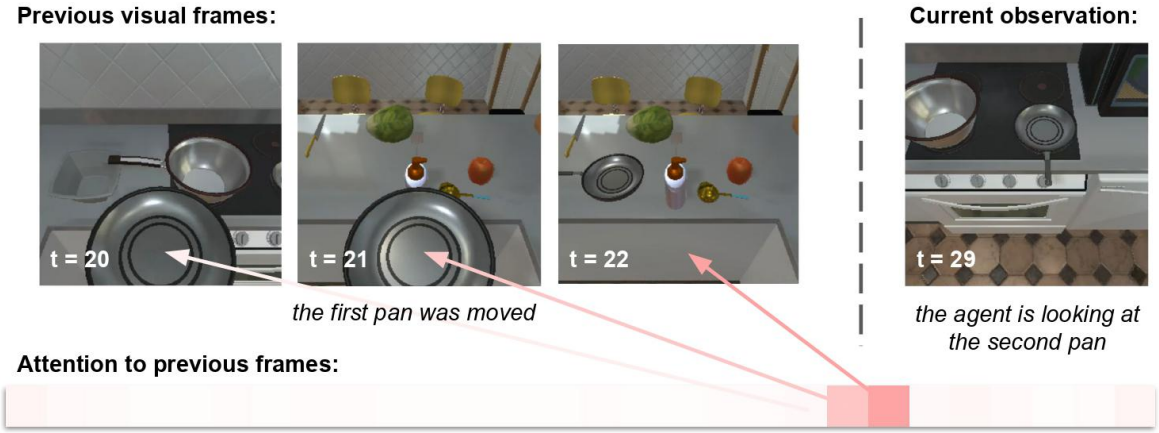
Goal: Grab an apple, cook it and put it in the sink. **Instructions:** Turn to your left twice so that you are facing the fridge. Open the fridge, grab an apple from the shelf and close the fridge door. *Walk to the left of the fridge to face the microwave.* Put the apple in the microwave and cook it for a few seconds before taking it back out and closing the microwave. Turn to face your left. Put the apple in the sink.

Figure A1: A visualization of normalized attention heatmap to previous visual observations, from white (no attention) to red (high attention). In this example, a microwave is first observed at the 8^{th} timestep, and is highlighted by the visual attention at the 19^{th} timestep when the agent is asked to put the apple in the microwave.



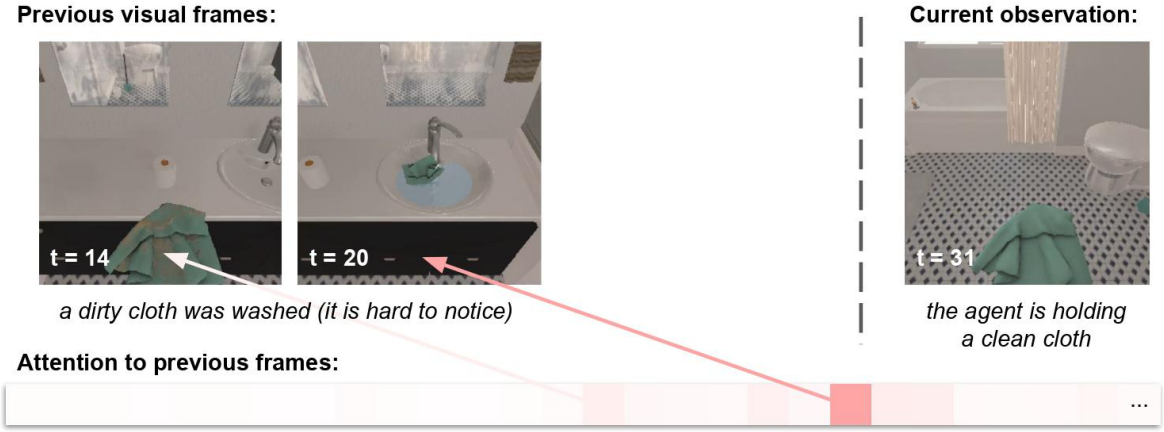
Goal: Prepare a potato to eat. **Instructions:** Walk to table, turn right to face counter next to sink. Pick up knife that is next to potato. Cut potato in half with knife. Place knife on top of salt shaker on the counter. *Pick up a middle slice of the potato.* Turn around and walk to the stove on the right. Face the stove. Place potato slice in the pot on the left side of the stove top. Pick up the pot from the stove that contains the potato, on the left hand side. Turn around and carry the pot with the potato to the table behind you. Place the pot on the table, on the edge of the plate.

Figure A2: A visualization of normalized attention heatmap to previous visual observations. In this example, the agent is asked to cut a potato (timesteps $17-18$) and to put a slice of it in a pot. At timestep 39 when the agent is asked to retrieve the sliced potato, it attends to frames at timesteps $17-18$ to decide where to go.



Goal: Move both pans from the stove top to the white counter top. **Instructions:** Turn right, walk towards the fridge and then turn left and go to the stove. Pick up the nearest pan on the stove top. Turn left and walk to the counter top. Put the pan down to the left of the white bottle. Turn back around and go back to the stove. *Pick up the other pan on the stove top.* Turn left and walk to the counter top. Put the pan down to the right of the apple.

Figure A3: A visualization of normalized attention heatmap to previous visual observations. In this example, the agent is asked to move two identical pans. It moves the first pan at timesteps $20 - 22$ and attends the frame at timestep 29 when moving the second pan (see the two corresponding pink squares).



Goal: Put a clean cloth in a drawer. **Instructions:** Go left to face the mirror over the counter. Pick the green cloth up from the counter. Move to the left and face the sink. Put the cloth in the sink and turn the water on and then off and pick the cloth up again. Move back and face the drawers under the counter and on the left. *Put the cloth in the bottom drawer and shut the drawer.*

Figure A4: A visualization of normalized attention heatmap to previous visual observations. In this example, the agent is asked to wash a cloth and to put it in a cupboard. The agent washes the cloth at timestep 20 but the washed cloth does not look very different from a dirty one. At timestep 31, the agent attends to the previous frames where the washing action is visible to keep track of the cloth state change.

of all heads and recursively multiplies attention weights of all transformer layers taking into account skip connections. Figures A1-A4 show examples where an E.T. model attends to previous visual frames to successfully solve a task. The frames attention weights are showed with a horizontal bar where frames corresponding to white squares have close to zero attention scores and frames corresponding to red squares have high attention scores. We do not include the attention score of the current frame as it is always significantly higher than scores for previous frames.

In Figure A1 the agent is asked to pick up an apple and to heat it using a microwave. The agent walks past a microwave at timestep 8, picks up an apple at timestep 18 and attends to the microwave frame in order to recall where to bring the apple. In Figure A2 the agent slices a potato at timesteps $17 - 18$ (hard to see on the visual observations). Later, the agent gets rid of the knife and follows the next instruction asking to pick up a potato slice. At timestep 39, the agent attends to the frames $17 - 18$ where the potato was sliced in order to come back to the slices and complete the task. In Figure A3 the agent needs to sequentially move two pans. While picking up the second pan at timestep 29, the agent attends to the frames $20 - 22$ where the first pan was replaced. In Figure A4 the agent is asked to wash a cloth and to put it to a drawer. The agent washes the cloth at timestep 20 but the cloth state change is hard to notice at the given frames. At timestep 31, the agent attends to the frame with an open tap in order to keep track of the cloth state change. To sum up, the qualitative analysis of the attention mechanism over previous visual frames shows that they are used by the agent to solve challenging tasks and aligns with the quantitative results presented in Section 4.3.

A.5. Visualizing language attention

Figure A5 illustrates transformer attention scores from an output action to input language tokens by comparing two models: (1) E.T. model trained from scratch, (2) E.T. model whose language encoder is pretrained as in Section 3.3. Similarly to the visual attention, we use attention rollout and highlight the words with high attention scores with red background color.

In the first example of Figure A5, the agent needs to pick up a bat. While the non-pretrained E.T. model has approximately equal attention scores for multiple tokens (those words are highlighted with pale pink color) and does not solve the task, the pretrained E.T. attends to "bat" tokens (highlighted with red) and successfully finds the bat. In the second example, the agent needs to first cool an egg in a fridge and to heat it in a microwave later. The non-pretrained E.T. has the similar attention scores for "microwave" and "refridgerator" tokens (they are highlighted with pink) and makes a mistake by choosing to heat the egg first. The pretrained E.T. agent has higher attention scores for the "refridgerator" tokens and correctly decides to cool the egg first. In the third example, the agent needs to pick up a knife to cut a potato later. The non-pretrained agent distributes its attention over many language tokens and picks up a fork which is incorrect. The pretrained E.T. agent strongly attends to the "knife" token and picks the knife up. The demonstrated examples show that the language pretraining of E.T. results in language attention that is better aligned with human interpretation.

A.6. Qualitative analysis

We show 3 successful and 2 failed examples of the E.T. agent solving tasks from the ALFRED validation fold. In Figure A6 the agent successfully heats an apple and puts it on a table. The agent understands the instruction *"bring the heated apple back to the table on the side"* and navigates back to its previous position. In Figure A7 the agent brings a washed plate to a fridge. The agent does not know where the plate is and walks along a counter checking several places. Finally, it finds the plate, washes it and brings it to the fridge. In Figure A8 the agent performs a sequence of 148 actions and successfully solves a task. This example shows that the agent is able to pick up small objects such as a knife and a tomato slice. The agent puts both of them to a plate and brings the plate to a fridge.

Among the most common failure cases are picking up wrong objects and mistakes during navigation. In Figure A9 the agent misunderstands the instruction *"pick up the bowl to the right of the statue on the table"* and decides to pick up a statue on the frame marked with red. It then brings the statue to a correct location but the full task is considered to be failed. Figure A10 shows a failure mode in an unseen environment. The agent is asked to pick up a basketball and to bring it to a lamp. The agent first wanders around a room but eventually picks up the basketball. It then fails to locate the lamp and finds itself staring into a mirror. The agent gives up on solving the task and decides to terminate the episode.

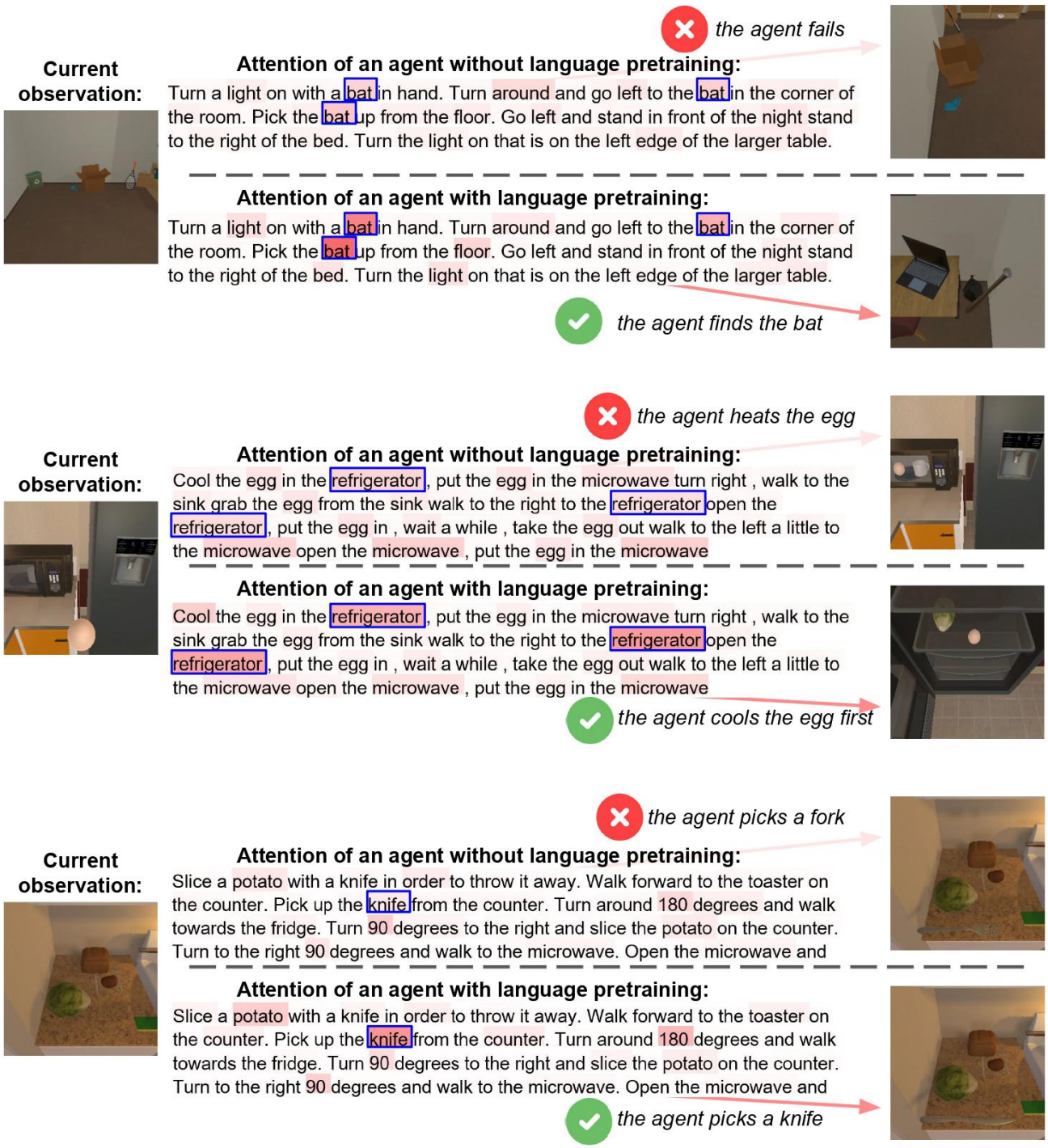
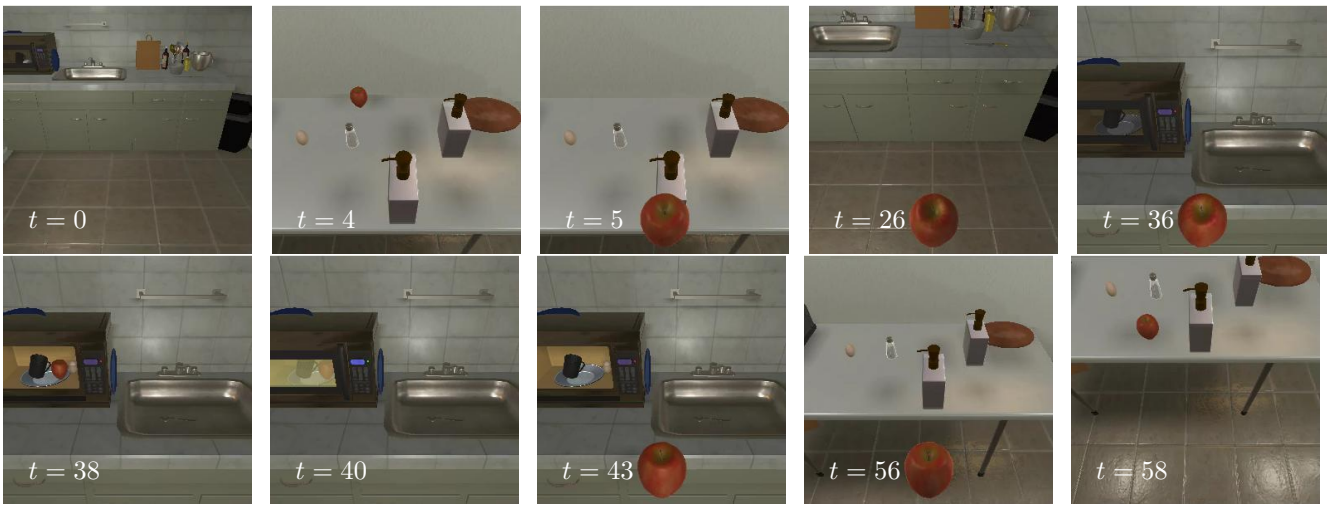
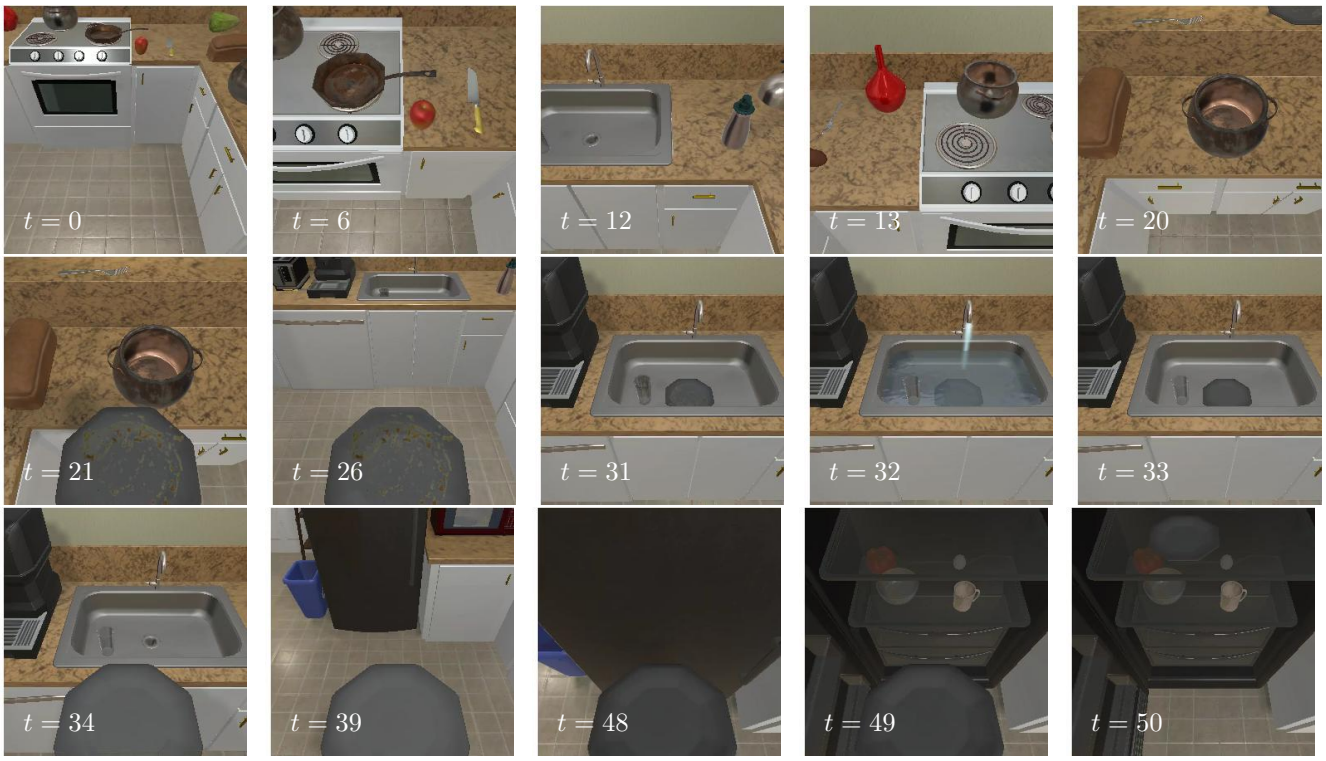


Figure A5: Visualizations of normalized language attention heatmaps, without and with the language encoder pretraining. Red indicates a higher attention score. We observe that the agent trained without language pretraining misses word tokens that are important for the task according to human interpretation (marked with blue rectangles). In contrast, the pretrained E.T. agent often is able to pay attention to those tokens and solve the tasks successfully.



Goal: Put a heated apple on the table. **Instructions**: Turn left and go to the table. Pick up the apple on the table. Go right and bring the apple to the microwave. Heat the apple in the microwave. *Bring the heated apple back to the table on the side.* Put the heated apple on the table in front of the salt.

Figure A6: Example of a successfully solved task. The agent picks up an apple, puts it into a microwave, closes it, turns it on, opens it, picks up the apple again, then navigates *back to the table on the side* and puts the apple on the same table.

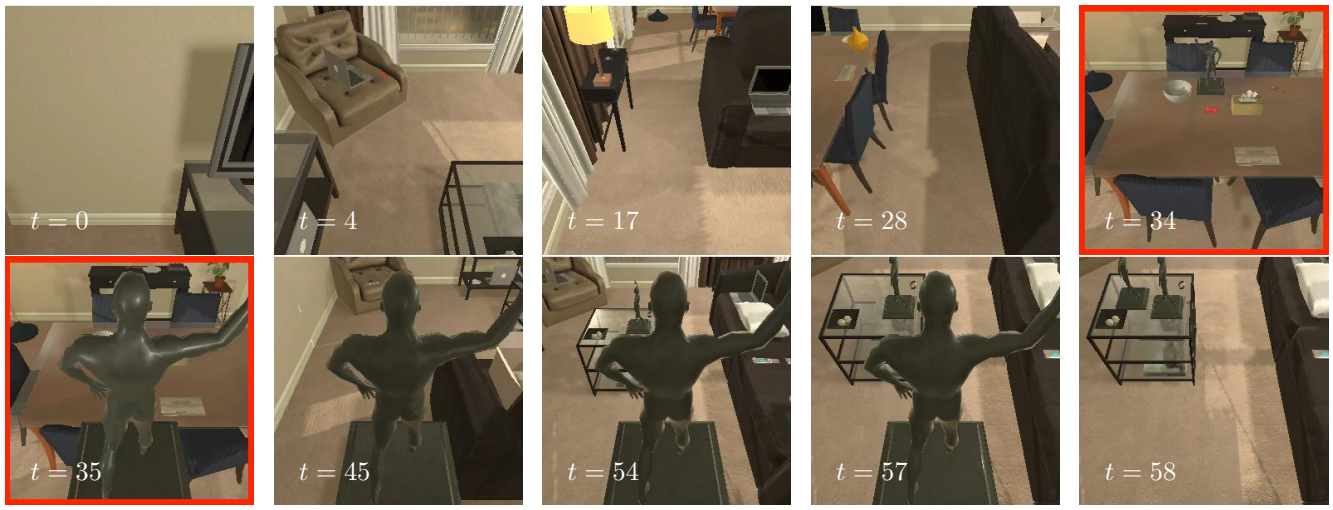


Goal: Place a rinsed plate in the fridge. **Instructions**: Walk ahead to the door, then turn left and take a step, then turn left and face the counter. Pick up the dirty plate on the counter. Walk left around the counter, and straight to the sink. Clean the plate in the sink. Turn left and walk to the fridge. Place the plate on the top shelf of the fridge. Place a pan containing slicing tomato in the refrigerator.

Figure A7: Example of a successfully solved task. The agent does not know where the dirty plate is and looks at several places on the counter (the first row). It then sees the plate in the corner of the top right image, picks it up, goes to a sink, opens a tap, picks the plate again, navigates to a fridge, opens it and puts the plate to the top shelf of the fridge.

Goal: Place a pan containing slicing tomato in the refrigerator. **Instructions**: Turn right, move to the table opposite the chair. Pick up the knife that is near the tomato. Turn left, move to the table opposite the chair. Slice the tomato that is on the table. Turn left, move to the counter that is left of the bread, right of the potato. Put the knife in the pan. Turn left, move to the table opposite the chair. Pick up a slice of tomato from the counter. Turn left, move to the counter that is left of the bread, right of the potato. Put the tomato slice in the pan. Pick up the pan from the counter. Turn left, move to in front of the refrigerator. Put the pan in the refrigerator.

Figure A8: Example of a successfully solved task. The agent uses 148 actions to complete the task. The agent picks up a knife from a table, slices a tomato in the first image of the second row, brings the knife to a stove, puts the knife on a plate, walks back to the table, grabs a tomato slice, returns to the stove, puts the tomato slice on the same plate, picks up the plate, navigates to a fridge, opens it, puts the plate with the knife and the tomato slice on a shelf and closes the fridge.



Goal: Move a bowl from the table to the coffee table. **Instructions**: Move across the room to the dining room table where the statue is. *Pick up the bowl to the right of the statue on the table.* Carry the bowl to the glass coffee table. Place the bowl on top of the coffee table between the statue and the square black tray.

Figure A9: Failure example in a seen environment. The agent correctly finds both dining and coffee tables but gets confused with *"the bowl to the right of the statue"* reference. The agent decides to pick up a statue instead of a bowl and fails to solve the task.

Goal: Look at a basketball in the lamp light. **Instructions**: Turn around and go to the foot of the bed. Pick up the basketball from the floor. *Turn around and go to the desk in the corner.* Turn on the lamp.

Figure A10: Failure example in an unseen environment. The agent is exposed to an unknown environment and fails to follow the navigation instructions. It wanders around the room, eventually finds a basketball but fails to locate a lamp and decides to terminate the episode in front of a mirror.