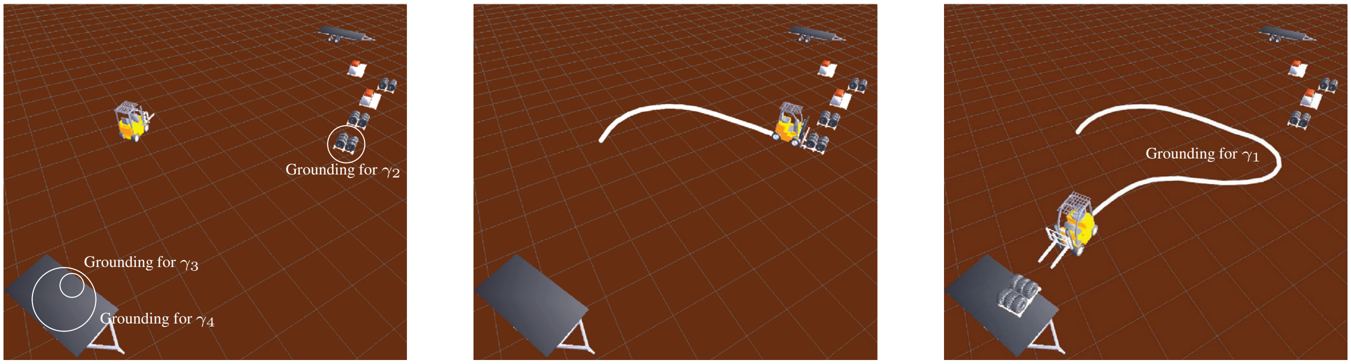

(a) Object groundings (b) Pick up the pallet (c) Put it on the truck

Figure 4: A sequence of the actions that the forklift takes in response to the command, "Put the tire pallet on the truck." (a) The search grounds objects and places in the world based on their initial positions. (b) The forklift executes the first action, picking up the pallet. (c) The forklift puts the pallet on the trailer.

of low-scoring examples were due to words that did not appear many times in the corpus.

For PLACE SDCs, the system often correctly classifies examples involving the relation "on," such as "on the trailer." However, the model often misclassifies PLACE SDCs that involve frame-of-reference. For example, "just to the right of the furthest skid of tires" requires the model to have features for "furthest" and the principal orientation of the "skid of tires" to reason about which location should be grounded to the language "to the right," or "between the pallets on the ground and the other trailer" requires reasoning about multiple objects and a PLACE SDC that has two arguments.

For EVENT SDCs, the model generally performs well on "pick up," "move," and "take" commands. The model correctly predicts commands such as "Lift pallet box," "Pick up the pallets of tires," and "Take the pallet of tires on the left side of the trailer." We incorrectly predict plans for commands like, "move back to your original spot," or "pull parallel to the skid next to it." The word "parallel" appeared in the corpus only twice, which was probably insufficient to learn a good model. "Move" had few good negative examples, since we did not have in the training set, to use as contrast, paths in which the forklift did not move.

4.3 End-to-end Evaluation

The fact that the model performs well at predicting the correspondence variable from annotated SDCs and groundings is promising but does not necessarily translate to good end-to-end performance when inferring groundings associated with a natural language command (as in Equation 1).

To evaluate end-to-end performance, we inferred plans given only commands from the test set and a starting location for the robot. We segmented commands containing multiple top-level SDCs into separate clauses, and utilized the system to infer a plan and a set of groundings for each clause. Plans were then simulated on a realistic, high-fidelity robot simulator from which we created a video of the robot's actions. We uploaded these videos to AMT, where subjects viewed the video paired with a command and reported their agreement with the statement, "The forklift in the video is executing the above spoken command" on a five-point Likert scale. We report command-video pairs as correct if the subjects agreed or strongly agreed with the statement, and incorrect if they were neutral, disagreed or strongly disagreed. We collected five annotator judgments for each command-video pair.

To validate our evaluation strategy, we conducted the evaluation using known correct and incorrect command-video pairs. In the first condition, subjects saw a command paired with the original video that a different subject watched when creating the command. In the second condition, the subject saw the command paired with random video that was not used to generate the original command. As expected, there was a large difference in performance in the two conditions, shown in Table 2. Despite the diverse and challenging language in our corpus, new annotators agree that commands in the corpus are consistent with the original video. These results show that language in the corpus is understandable by a different annotator.

	Precision
Command with original video	0.91 (±0.01)
Command with random video	0.11 (±0.02)

Table 2: The fraction of end-to-end commands considered correct by our annotators for known correct and incorrect videos. We show the 95% confidence intervals in parentheses.

We then evaluated our system by considering three different configurations. Serving as a baseline, the first consisted of ground truth SDCs and a random probability distribution, resulting in a constrained search over a random cost function. The second configuration involved ground truth SDCs and our learned distribution, and the third consisted of automatically extracted SDCs with our learned distribution.

Due to the overhead of the end-to-end evaluation, we con-