We thank all reviewers for their constructive feedback! We are encouraged that they found our contribution interesting (R4), addressing a hard problem that is important in robotics (R4), while also extremely relevant to NeurIPS (R1) in the area of robot control from language and vision (R2). Our extensive experiments (R4) demonstrate impressive performance (R1), show the relevance of each component (R3, R4), and that it is outperforming previous methods on simulated robotic tasks (R1, R3) in more complex environments (R4). Moreover, reviewer R4 highlighted our model's ability to modulate actions via descriptors like "little" or "all". **In our supplemental material, we provide the full source-code for training and testing, a description and examples of verbal commands by human teachers, and robot videos – please consult index.html**. The primary concern raised regards more detailed explanations of component integration (R1), human evaluation (R2), FRCNN/Glove choices (R2), potential use of simpler methods (R3) and modularity (R4). We appreciate the feedback and commit to using the extra 1-page allowed for the camera-ready version of accepted papers to substantially expand explanations and address this, as well as add requested references.

R1,R2 **More details on human data collection?** We had five annotators watch pre-recorded agent trajectories and verbally describe the action the robot was performing in unconstrained language. Annotators were graduate students familiar with robotics but not the aim of this project. We collected 200 such descriptions (40 per annotator). Descriptions were transcribed and manually templated by marking replaceable noun-phrases and adverbs. We automatically generated training descriptions by filling these slots with synonyms based on the scene (line 233-239) – **see supplemental material**. This resulted in 99864 unique task descriptions randomly reduced to 45k. In experiment "Generalization to New Users", users typed an instruction and saw the result in a physics-based simulation in real-time (line 263).

R2 **Faster R-CNN isn't trained on data that looks anything like this.** We fine-tuned/pre-trained the FRCNN model (from ResNet-101 trained on COCO) on 40k arbitrarily generated environments of our simulator. After fine-tuning, the certainty on FRCNN on our objects is above 98%. FRCNN was chosen because it is a commonly-used method with reasonably high performance and largely understood pros and cons. We do not claim contributions to object detection, but rather to the integration of language and vision for robot control. While we are open to using simpler approaches, FPFH would not be applicable since it is a 3D point-cloud approach requiring access to a depth camera.

R1, R3 **How are all components integrated? Simpler components?** The components of our model, explained in section 3.2 and 3.3, are integrated sequentially. After pre-processing of the image and language data (section 3.1), our model takes the input data and converts them into a task-specific embedding within the semantic model (section 3.2). In turn, this embedding is then used to generate the hyper-parameters of the low-level controller (line 152), namely weights, current phase, and desired phase progression (line 173). With these parameters, we define a motor primitive determining the next robot motion (line 186) as well as the entire trajectory. Especially the latter is a benefit of our approach over simpler FF networks. We commit to substantially expanding section 3.4 (R1) in the camera-ready version as well as adding a discussion explaining why simpler models did not work as well (R3).

R2 **Why is PayAttention performing poorly?** We used the original code from the PayAttention paper and have repeatedly sought out the assistance of the main author (who thankfully provided substantial support) to ensure that our usage exactly follows the method and protocol required. Regarding performance, we argue that in our case, adverbs and adjectives play a critical role in disambiguating objects and modulating the behavior. PayAttention, however, primarily focuses on objects that can be clearly differentiated by their noun.

R4 **Why end-to-end and not a multi-staged approach?** As the reviewer pointed out, multi-staging requires a significant amount of additional (human) feature engineering, which would, at the same time, limit our approach to these features and may also be fragile in terms of generalizing to new words. Our framework learns how language affects the behavior (type, goal position, velocity, etc.) automatically, while also learning the control itself. Another advantage of end-to-end is that the overall system can be trained such that the individual components harmonize. This is particularly important for the interplay of language embedding and control when using language as a modifier for trajectory behaviors.

R2 **We know deep contextual models like BERT can generally better process sequences.** This is a great suggestion; our approach easily allows for the integration of alternative approaches for language, e.g., BERT. GloVe was chosen due to its simplicity, but to demonstrate our framework's extensibility, we incorporated BERT as suggested. Due to time constraints, we trained only a single model **which achieved 97% on the picking and 65% on the pouring task**, which is comparable to our original model with 98% and 85% respectively. Better performance can likely be achieved with more careful integration and tuning of the parameters. Still, we argue that this ability to replace language models with more recent SotA methods quickly is an appealing feature of our approach that allows steady improvements.

R2 **The method requires strong supervision (language to object referents) during training, which requires backing off to templates to get enough training data to use an RL method.** We would like to clarify: our approach is a pure imitation learning technique, and no RL is involved. RL could potentially be an option, but learning policies that bridge language, vision, and control with a limited amount of trials (< 100K) would be extremely challenging.