We thank the reviewers for their insightful comments. We are glad reviewers acknowledge our promising approach of using hierarchical control and relational information to generalize bimanual manipulation imitation learning **R1, R2, R3, R4**. We are encouraged by comments about the soundness of our experiments and theory **R1, R3**. We will fix the minor issues and typos in the updated version. Below we clarify our approach and address specific concerns.

R1, R2 *Related work and baselines* To the best of our knowledge, using deep imitation learning for bimanual manipulation have only involved MLP [1] and RNN [2]. [4] uses graph representation for visual imitation learning. MLP did not suit our task as we wanted to predict trajectory sequences. The GRU-GRU model presented in Table 1 was an implementation of RNN, which we adapted to our simulation environment. Graph neural networks with skip connections in the context of bimanual manipulation have not been previously investigated. We also experimented with GAIL [3] but it did not perform well. Errors for the first primitive were at least 6 times higher than HDR-IL. This is likely due to the low variances of our expert demonstrations, making the policy distribution difficult to fit.

R1, R4 *29% accuracy rather low...* The success rate is low because the task is intended to test the limits. In our high-precision table lifting task, we showed 100% success rate in Table 1. This demonstrates our HDR-IL model greatly improved generalization compared to prior methods. To deepen our investigation, we designed the more challenging peg-in-hole task to test multiple generalizations in far out (up to 130) time steps.

R1 *...an error of 5cm with HDR-IL seems unreasonably high.* The Euclidean distance errors shown in Table 1 **are normalized** to per datapoint position. These errors do not necessarily reflect the simulation task performance. Task success is largely determined by the accuracy at the end of a primitive. Errors in our model tend to be at the beginning as suggested by the trajectories in Figure 4. Furthermore, Table 2 of the Appendix show the 5cm error is largely driven by the Extend primitive, which is less crucial for the task success. The Extend primitive has the largest step sizes between datapoints, as illustrated in Figure 3(a). The bigger steps leads to bigger prediction errors.

R1 *...actual pose of one arm...relative pose of the second arm...* We ran tests and found using relational data does not generalize as well as graph structure with absolute poses. Euclidean distances were 8.15 ± 4.71 with success rate 16% compared to 5.64 ± 5.17 and 72% for ResInt in Table 1. We will add these to our baseline in an updated version.

R1 *...effect of orientation...* Orientation of the grippers were included in the prediction to improve inverse kinematics accuracy. We will quantify orientation errors of the gripper quaternions in an updated version. The generalization of the model to different table starting orientations is a topic for future study.

R1 *Not clear if...specific to bimanual manipulation* Our method is NOT specific to bimanual manipulation, but the complex dependencies in the bimanual manipulation setting highlight the value of graphs in our experiments.

R2 *...why did the ground truth demonstration fail...* The failures in demonstration are due to approximations of the inverse kinematics (IK) solver. Such failure cases are rare (<0.7% of 2500 demonstrations). The IK solver sometimes outputs unusual movements between poses which cause the table to be dropped. The arm trajectory in the demonstration will change accordingly. Our model can account for these uncertainties in demonstrations because of the stochastic sampling design shown in Figure 2.

R2 *...limited to the provision of the graph structure.* Providing the graph structure is a common way to encode inductive bias e.g. [4], not a limiting factor. Learning the graph structure is related to model selection and causal reasoning, which is out of scope of this paper. However, we did experiment with a few different graph structures before arriving at our fully connected graph with attention.

R2 *...does not show off how complex two handed settings...* Our simulations include complex **physical contacts** and **friction**. We did not use a flexible table. The peg-in-hole task was specifically designed to increase the task complexity: assembling the table before lifting requires more coordination between the two arms. We showed a failed attempt at lifting the table using one arm in our supplementary videos. Doing the same for either piece in the peg-in-hole task will have the same result. A hammer and nail task would be an even more complicated experiment for future work.

R2 *...manually crafted primitives...contradicting to learning...* This is NOT a contradiction. The primitives are constructed based on the control commands executed in simulation. During training, the primitives are labeled. Our model learns to compose the primitives for high level planning. It also learns the low level dynamics for each primitive. During inference, the model predicts the primitives based on the demonstration trajectories.

[1] R. Chitnis, et al., "Efficient Bimanual Manipulation Using Learned Task Schemas," *ICRA*, 2020

[2] D. Rakita, et al., "Shared Control–Based Bimanual Robot Manipulation," *Science Robotics*, 2019

[3] J. Ho, et al., "Generative Adversarial Imitation Learning," *NeurIPS*, 2016

[4] M. Sieb, et al., "Graph-structured Visual Imitation," *CoRL*, 2020