AugInsert: Learning Robust Visual-Force Policies via Data Augmentation for Object Assembly Tasks

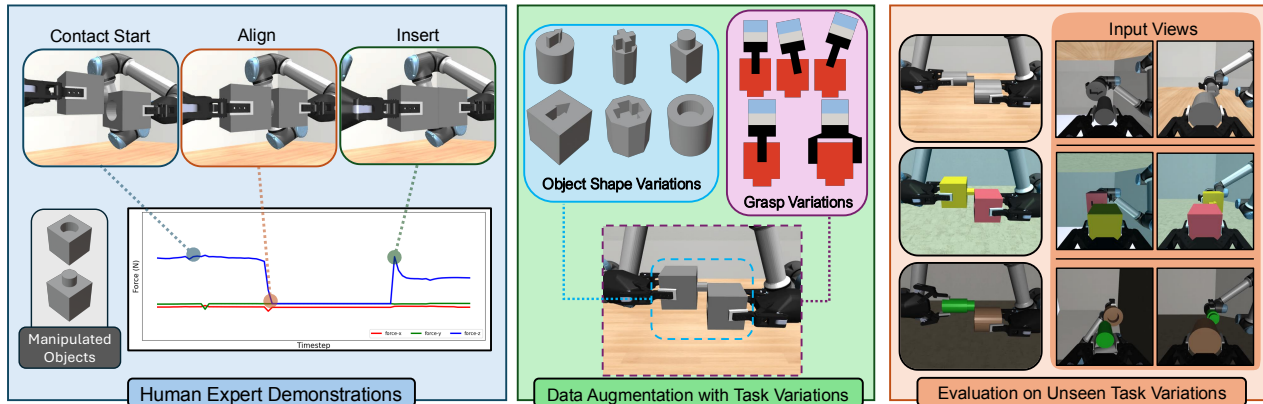Ryan Diaz[1], Adam Imdieke[1], Vivek Veeriah[2], Karthik Desingh[1]

Fig. 1: AugInsert is a data collection and policy evaluation pipeline aimed towards analyzing the robustness of a multisensory (vision, force-torque, and proprioception) model with respect to different observation-level task variations in object shape, grasp pose, and visual environmental appearance. Our framework introduces task variations to a dataset of human-collected demonstrations through a system of online data augmentation.

Abstract— **Operating in unstructured environments like households requires robotic policies that are robust to out-of-distribution conditions. Although much work has been done in evaluating robustness for visuomotor policies, the robustness evaluation of a multisensory approach that includes force-torque sensing remains largely unexplored. This work introduces a novel, factor-based evaluation framework with the goal of assessing the robustness of multisensory policies in a peg-in-hole assembly task. To this end, we develop a multisensory policy framework utilizing the Perceiver IO architecture to learn the task. We investigate which factors pose the greatest generalization challenges in object assembly and explore a simple multisensory data augmentation technique to enhance out-of-distribution performance. We provide a simulation environment enabling controlled evaluation of these factors. Our results reveal that multisensory variations such as *Grasp Pose* present the most significant challenges for robustness, and naive unisensory data augmentation applied independently to each sensory modality proves insufficient to overcome them. Additionally, we find force-torque sensing to be the most informative modality for our contact-rich assembly task, with vision being the least informative. For additional experiments and qualitative results, we refer to the project webpage https://bit.ly/47skWXH.**

I. INTRODUCTION

Robust manipulation in unstructured environments requires robots to adapt to unforeseen variations in object properties, positions, and environmental conditions. This is particularly challenging for contact-rich tasks, where physical interaction plays a crucial role in successful execution. We focus on the peg-in-hole assembly task as a representative

example of contact-rich manipulation. Like inserting a K-cup pod into a coffee machine, capping a bottle, or plugging in a cable, the peg-in-hole task requires precise control and adaptation to variations in contact forces and object geometries throughout the insertion process. These shared characteristics make the peg-in-hole task a valuable benchmark for studying and improving the robustness of multisensory policies for a broader range of contact-rich manipulation tasks.

In this paper, we address the challenge of robust manipulation in contact-rich environments by developing a multisensory policy learning framework that includes force-torque sensing. Our pipeline processes multiple camera views and force-torque (F/T) readings from a dual-arm setup. In an effort to increase robustness, we also explore a multisensory data augmentation method via trajectory replay that can introduce both sensor-specific (camera and F/T sensor) variations as well as physical factors such as manipulated object shape, peg and hole geometries, and grasp pose variations that affect the sensing modalities. Through this approach, we can expand small expert datasets to learn robust manipulation policies that can handle a wide variety of environmental conditions.

To analyze the robustness of our model with respect to specific observation-level task variations and understand the effect of our data augmentation method, we develop an experimental setup in the MuJoCo [1] simulation environment with a dual-arm robot that can manipulate objects with peg and hole geometries to complete the assembly task. Our experiments show that certain variations, such as *Grasp Pose* variations, cause large drops in success rate for our task and so should be included in training data through data augmentation in order to ensure robustness to

[1]R. Diaz, A. Imdieke, and K. Desingh are with the Department of Computer Science and Engineering at the University of Minnesota, Twin Cities. {diaz0329, imdie022, kdesingh}@umn.edu

[2]V. Veeriah is with Google DeepMind. vveeriah@google.com

these variations. Additionally, we conduct ablation studies to understand how each sensory modality in the multisensory setup affects the performance of the contact-rich assembly task. These studies also help identify the specific modalities impacted by each variation. We observe that touch provides the most relevant information for the task and supports model robustness; visual input, on the other hand, has the least significant impact on generalization ability while also being susceptible to many of our task variations. We provide an extensive discussion of these results in the following sections.

A. Contributions

The main contributions of our work are

- A dual-arm object assembly task formulation along with a set of physical and sensor-based task variations that introduce a variety of perturbations to the environment state.
- A multisensory vision and force-torque policy learning framework based on the Perceiver IO [2] architecture to perform this contact-rich task.
- A publicly available simulation environment for our dual-arm object assembly that supports physical and sensor-based task variations, enabling standardized benchmarking and evaluation.
- Extensive experiments analyzing the most challenging task variations for generalization in our assembly task, along with an exploration of data augmentation strategies to enhance robustness.

II. RELATED WORK

A. Multisensory Contact-Rich Manipulation

Multisensory policy learning using vision and F/T data for contact-rich peg-in-hole insertion has been widely studied. Lee et al. [3], [4] developed a self-supervised learning method to learn a multisensory representation using vision and F/T that can transfer across different peg and hole shapes, and Wu et al. [5] created a reward learning framework for the peg-in-hole task based on task progress. More recently, Spector et al. [6], [7] developed a multiview and multisensory system for localizing and performing realistic insertion tasks, Chen et al. [8] used a transformer [9] encoder for vision and F/T inputs to learn a higher-quality representation, and Kohler et al. [10] leveraged symmetry in the peg-in-hole task by using equivariant networks to improve sample efficiency. Although these works can achieve efficient peg-in-hole assembly, their generalization studies are limited in scope when evaluating robustness to both physical and sensory task variations. Our work aims to determine the types of variations that present the greatest challenges for generalization in this task in order to focus our efforts in increasing robustness.

B. Evaluating Generalization Abilities of Learned Policies

Generalization is difficult to define in robot manipulation policy learning as there are several factors in the robot's environment that could vary from training phase to the evaluation phase. There have been recent efforts to perform in-depth analyses of the generalization abilities of visuomotor robotic policies by decomposing task environment variations into individual variation "factors" [11]–[14] to isolate different types of perturbations to the environment state. We aim to bring this type of analysis to the multisensory domain by introducing a set of variation factors that perturb F/T and proprioceptive inputs in addition to image inputs.

C. Data Augmentation for Increased Model Robustness

Training on diverse demonstration datasets has allowed for the development of robot policies that can accomplish complex manipulation tasks while being robust to certain environmental variations, and a large focus has been placed on curating these datasets with minimal human intervention [15]–[17]. A common technique to enhance a model's robustness without requiring additional human effort in data collection is data augmentation, which involves transforming input data while preserving the original labels. This approach is most commonly used with image inputs, where transformations such as cropping, flipping, and color adjustments are applied to help the model learn invariance to these changes [6], [18]–[20]. Image augmentation can also be performed at the semantic level using generative models [21]–[23]. However, applying this type of augmentation to contact-rich tasks poses challenges, as these tasks involve *physical* variations (e.g., object size, shape), which may introduce non-independent perturbations across the multisensory input that cannot be captured through conventional offline augmentation. Our work explores an online augmentation method that can include multisensory variations in the training dataset.

III. TASK SETUP

A. Assembly Task Definition

Our experimental setup consists of a dual-arm robot manipulator with a multisensory configuration, featuring two F/T sensors and two RGB cameras attached to its wrists. The robot is tasked with performing an insertion assembly, where one arm's gripper holds a peg-shaped object and inserts it into a hole-shaped object held by the other arm's gripper. Since our focus is on the contact phase of the assembly, the objects are already in contact at the start of the task.

The objective of the robot learning framework is to execute the assembly task without explicit information about the object geometries or peg-and-hole shapes, while maintaining robustness to various task variations. We take the behavior cloning approach where expert demonstrations are used to clone the contact-rich manipulation policy to perform the assembly.

B. Task Initialization

The task is initialized with the peg and hole offset within a range of [1.5cm, 3.0cm] along both the X and Y axes relative to the object coordinate frame (perpendicular to the direction of insertion). Our setup ensures that while the arm holding the peg moves, the other arm remains compliant, applying a constant force until the peg and hole are aligned. To define a successful task rollout, we consider position coordinates
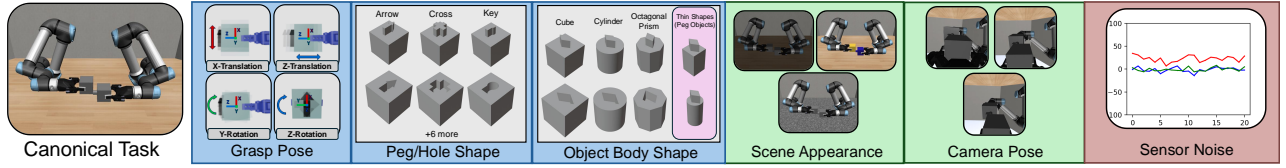
Fig. 2: A visualization of each of the task variations used in our environment setup. We differentiate between physical task variations (in blue) and sensor-based task variations that target vision (green) and force-torque proprioception (red).

$\mathbf{p} = (x_p, y_p, z_p)$ for the peg object and $\mathbf{h} = (x_h, y_h, z_h)$ for the hole object in the global coordinate frame. We set thresholds $\mathbf{d} = (d_x, d_y, d_z)$ such that during a successful insertion, $|\mathbf{p}_i - \mathbf{h}_i| < \mathbf{d}_i$ for all axes $i \in \{x, y, z\}$.

C. Task Variations

To evaluate the robustness of our trained models, we design a set of observation-level task variations which alter the distribution of incoming observations while preserving the underlying task. Additional details on these variations can be found in supplementary material. In total, there are six variations that are part of the experiments (see Figure 2 for sample visualizations of these variations):

1) **Peg and Hole Shape:** There are 9 possible peg and hole shapes, with each peg and hole sharing the same shape and allowing for a tolerance to ensure insertion compatibility.

2) **Object Body Shape:** There are 3 possible object body shapes, and the peg and hole in a given pair may not share the same shape. Additionally, we create thinner versions of the peg to introduce variability, resulting in 6 total peg and hole object pairs.

3) **Grasp Pose:** Our grasp pose variation follows the approach in [24], which provides a more detailed overview.

4) **Scene Appearance:** This category encompasses variations in lighting, floor texture, and object color.

5) **Camera Pose:** We slightly perturb the position and orientation of the wrist cameras between demonstrations, while keeping them constant within each demonstration.

6) **Sensor Noise:** We add zero-mean Gaussian noise to low-dimensional measurements, including both force-torque and proprioception readings for both arms.

Canonical Task Setup: We define a "canonical" task setup which represents an environment without any task variations applied. For discrete task variations, we choose the `key` peg and hole shape, `cube` object body shape, and `light-wood` floor texture in our canonical setup in simulation.

IV. METHODOLOGY

A. Imitation Learning Framework

Observation and Action Spaces: In our task, the observation space is defined as a composition of four modality spaces, $\mathcal{O} = \mathcal{I}_{left} \times \mathcal{I}_{right} \times \mathcal{T} \times \mathcal{S}$. The image spaces, $\mathcal{I}_{left} \subseteq \mathbb{R}^{84 \times 84 \times 3}$ and $\mathcal{I}_{right} \subseteq \mathbb{R}^{84 \times 84 \times 3}$, represent 84×84 RGB wrist views from the left and right arms. The tactile space, $\mathcal{T} \subseteq \mathbb{R}^{32 \times 12}$, corresponds to a history of the last 32 force and torque readings from both arms (concatenated),

while the robot state space, $\mathcal{S} \subseteq \mathbb{R}^{14}$, represents the end-effector positions and orientations (expressed as quaternions) for both arms (concatenated). Our action space, $\mathcal{A} = [0, 1]^3$, consists of end-effector position deltas relative to the current pose.

Policy Learning: The goal is to learn a policy $\pi : \mathcal{O} \to \mathcal{A}$ that maps observations to actions, enabling task completion. In the imitation learning setting, an expert policy π^* is provided, where $a^* = \pi^*(o)$ represents the optimal action for an observation $o \in \mathcal{O}$. Our objective is to learn a policy π that closely resembles the behavior of π^*. There are several approaches to learning such a policy from demonstrations, with the simplest being *behavior cloning*. In behavior cloning, the expert provides a dataset of N demonstration trajectories $\mathcal{D} = \{\{(o_i, a_i^*)\}_{i=1}^{n_j}\}_{j=1}^{N}$, where n_j is the horizon for demonstration j. The policy π is then trained to replicate the expert actions from π^* for the corresponding observations using supervised learning. Our observation encoder and policy network (shown in Figure 3) are trained end-to-end using an L_2 loss between expert and predicted actions.

B. Data Collection with Human Experts

We collect a dataset of 50 human demonstrations in our simulation environment, built using the Robosuite framework [27] with MuJoCo [1] as the simulation engine. All demonstrations are performed in the canonical environment setup described in Section III-C. A human expert teleoperates the robot's moving arm via keyboard inputs, with actions recorded as the difference between the end-effector positions in consecutive frames. The simulation automatically terminates and records the demonstration upon detecting a successful completion.

C. Multisensory Data Augmentation

We hypothesize that *offline* data augmentation methods (such as those used for image augmentation) may not be effective for increasing robustness to multisensory variations. To address this, we explore an *online* data augmentation technique by replaying human-generated trajectories on task instances with identical initial offsets but with a subset of task variations applied. Given the previously-defined dataset of expert demonstration trajectories $\mathcal{D}$, task variations $\mathcal{V}$ (e.g. *Grasp Pose*, *Peg/Hole Shape*, etc.), and a function $f_{\mathcal{K}} : \mathcal{O} \to \mathcal{O}$ that returns an input observation with a subset $\mathcal{K} \subseteq \mathcal{V}$ of task variations applied, our online augmentation process takes an expert demonstration $d_j = \{(o_i, a_i^*)\}_{i=1}^{n_j} \in \mathcal{D}$ and outputs a set of new demonstrations
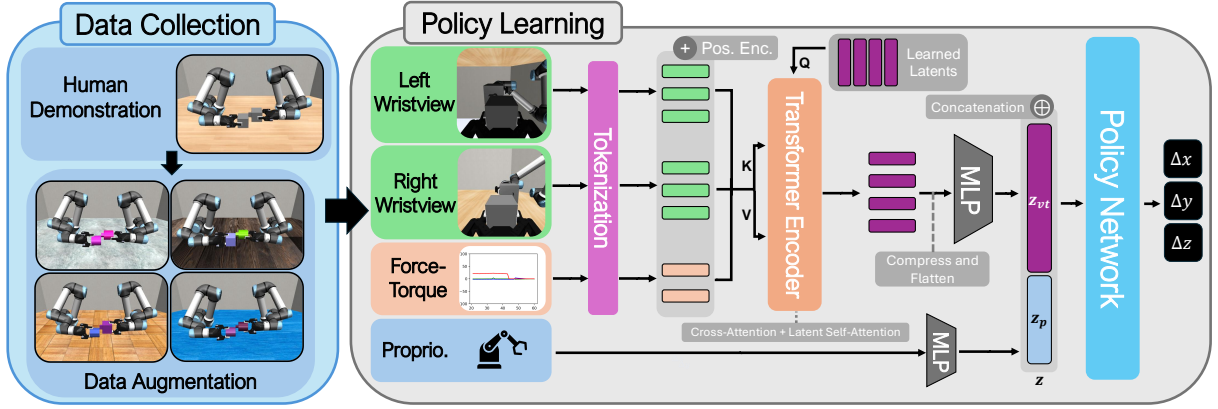
Fig. 3: An overview of our data collection and policy learning framework. We use BC-MLP [25] with a multilayer perceptron policy network to output actions. Image and force-torque observations are encoded with a visuotactile transformer [8] that includes a cross-attention step with a set of learned latent vectors (similar to Perceiver IO [2], [26]). More details on our network architecture can be found in our supplementary material and website.

$\Omega_{d_j} = \{\{(f_{\mathcal{K}}^t(o_i), a_i^*)\}_{i=1}^{n_j}\}_{t=1}^T$, where T is the number of augmentations per expert demonstration. The indexed functions $f_{\mathcal{K}}^1, \ldots, f_{\mathcal{K}}^T$ indicate that although each application of task variations $\mathcal{K}$ is different between each generated demonstration in Ω_{d_j}, the specific application of $\mathcal{K}$ is consistent for each observation in a given augmented demonstration. After the augmentation process, we construct a new dataset $\hat{\mathcal{D}} = \mathcal{D} \cup \left(\bigcup_{n=1}^N \Omega_{d_n} \right)$ that can be used for training. It can be noted that in simulation, this data augmentation approach resembles domain randomization [28], [29] by randomizing the simulator state, but our focus here is on variations for a multisensory setup as opposed to a camera-only one.

V. EXPERIMENTAL SETUP

We conduct experiments to evaluate the robustness of our model with respect to each of the 6 implemented task variations (see Section III-C), as well as a combination of all variations (denoted as *All Variations*).

A. Training and Evaluation Details

In all experiments, we train the models for 25000 training steps total. In simulation, we draw inspiration from [25] and perform 50 rollouts every 1250 training steps on the same task variations used in the training dataset. The model checkpoint that achieves the highest success rate in these rollouts during training is selected for evaluation on unseen scenarios. This training process is conducted over 6 random seeds per model, and the performance is averaged across all seeds during evaluation. A rollout is considered successful if it results in a successful insertion, and it is deemed failed if the maximum horizon is exceeded without insertion. Additionally, a rollout fails if the force-torque measurement surpasses a predefined threshold (100N of force or 6N-m of torque in simulation), to prevent unsafe behavior that could damage the robot arms or objects.

B. Evaluation on Unseen Task Variation Instances

To ensure rigorous evaluation, we explicitly separate task variation instances encountered during training from those used for evaluation within each variation category.

For instance, when training with *Grasp Pose* variations in simulation, we include demonstrations involving x-axis translation and z-axis translation and rotation, while reserving y-axis rotation for evaluation. This approach guarantees that the model encounters unseen variations during evaluation, enabling us to assess its generalization to out-of-distribution inputs across all variation categories. A detailed overview of training versus evaluation instances for each variation category is provided in Table I.

Task Variation	Train Instances	Eval Instances
Grasp Pose [24]	XT, ZT, ZR	XT, ZT, ZR, YR
Peg/Hole Shape	key, circle, cross	arrow, u, pentagon, line, hexagon, diamond
Object Body Shape	**Peg/Hole Objects:** cube, cylinder	**Hole Object**: cube, cylinder, octagonal prism, **Peg Object**: thin cube, cylinder, and octagonal prism
Scene Appearance	6 floor textures, object color	14 unseen floor textures, object color, lighting

TABLE I: Instances for task variations during training (if included the training set) and evaluation. Task variations not in this table are the same both in training and evaluation.

VI. EXPERIMENTS AND RESULTS

In this section, we address a series of questions related to generalization through experimental evaluations.

A. Which task variations are most difficult to generalize to?

In determining the difficulty of generalizing to each of our task variations, we train a model exclusively on human-collected demonstrations without any task variations applied and report the success rate for each task variation during evaluation. These results can be found in Figure 4.

Takeaways: We observe that *Grasp Pose* variations the hardest challenge to generalize to out of all of the individual task variations, as we see a drop from a mean success rate of
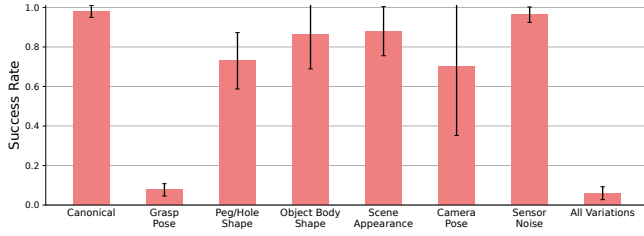
Fig. 4: Success rates on each task variation for a model trained exclusively on non-augmented human demonstration data. Error bars represent one standard deviation from the mean. The model suffers the largest success rate drop compared to the canonical environment when evaluated on *Grasp Pose* variations, and subsequently does poorly when evaluated on *All Variations*.
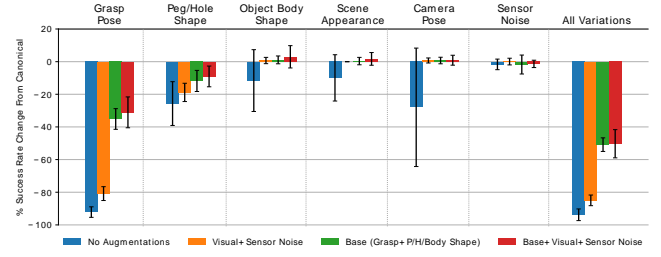

Fig. 5: % success rate changes on each task variation for models trained on different subsets of task variations. Error bars represent one standard deviation from the mean. The addition of grasp variations to the training set greatly improves generalization to unseen grasp variations, while visual variations and sensor noise do not have any significant effect on generalization ability on physical task variations.

0.987 on *Canonical* rollouts with no task variations to **0.087** when *Grasp Pose* variations are applied. We hypothesize that the large negative impact on performance comes from the significant perturbation that grasp variations apply to all sensing modalities, unlike other variations such as *Scene Appearance* and *Sensor Noise* which only target specific modalities and thus have a smaller negative impact for the overall model.

B. Which task variations included in the training set produce the largest impact on robustness?

To evaluate the effect of introducing task variations to the training dataset, we evaluate models trained on datasets augmented with different subsets of our task variations. These datasets contain the original collected human demonstrations as well as 6 augmentations per demonstration, with each augmentation containing a composition of all the task variations included in the dataset (which we refer to as the "training set variations" for that specific dataset). We evaluate these models both on instances of their training set variations that were unseen during training as well as all variations not included in the training set (which we refer to as the "evaluation set variations" for that specific dataset). In Figure 5 we report % success rate change from the canonical environment success rate averaged over 6 seeds for each variation, defined as

$$\% \text{ success rate change} = \frac{\text{task var. success} - \text{canon. success}}{\text{canon. success}}$$

Takeaways: We observe that including only unisensory variations such as *Scene Appearance* (visual variations) and *Sensor Noise* do not significantly improve policy robustness with respect to multisensory variations such as *Grasp Pose*, signaling that unisensory data augmentation cannot handle multisensory perturbations. Although there is a limited amount of visual variations in this dataset, supplemental experiments showed that performance was not improved even with a greatly expanded set of visual and sensor noise variations. On the other hand, we observe that the additions of *Peg/Hole Shape*, *Object Body Shape*, and *Grasp Pose* (considered the `Base` variations for this task) to the training set greatly reduce the generalization gap on unseen instances of these variations during evaluation. Curiously, we also

observe a reduced generalization gap for a dataset with the `Base` training set variations on the evaluation set variations of *Scene Appearance* and *Camera Pose*, even though these variations had not been explicitly included in the dataset. This may be due to the similarity between the effects of applying grasp variations and perturbing the camera pose, as both alter the view of the object held by the gripper and the opposing object. Additionally, the resulting visual variations may have contributed to improving the model's robustness to changes in scene appearance. Explicitly adding both visual variations (*Scene Appearance* and *Camera Pose*) and *Sensor Noise* to the training set does not further enhance generalization to their respective variations during evaluation.

C. Can increasing the number of augmentations per demonstration improve robustness?

Building off of our investigation into determining the ideal training set variations, we also seek to analyze the effect of adding more augmentations per demonstration involving these variations. Aligning with the previous experiment, we choose *Grasp Pose*, *Peg/Hole Shape*, and *Object Body Shape* as our training set variations, and train models on datasets with different numbers of augmentations per human demonstration. Success rates on each task variation are reported in Figure 6.
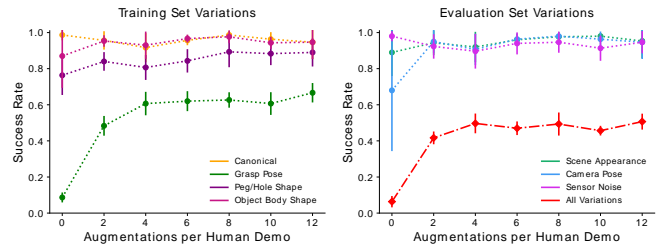

Fig. 6: Success rates on each task variation for models trained on a base set of training variations with different number of augmentations for each human demonstration. *All Variations* represents a composition of both training set and evaluation set variations. Error bars represent one standard deviation from the mean. Success rate on *Grasp Pose* variations increases the most with an increasing number of augmentations, while the other task variations maintain stable success rates.

Takeaways: The most significant improvement in success

rate as the number of augmentations increases seems to be the performance on *Grasp Pose* evaluations, with a more subtle upward trend in the other training set variations. Since the task variations of *Peg/Hole Shape* and *Object Body Shape* are discrete variations with a small subset of all possible shapes being included in the training set, they would benefit less from having more augmentations as the dataset would start to contain redundant instances of these variations. *Grasp Pose* variations, on the other hand, are continuous and so would benefit more from a larger sample of grasps. For the evaluation set variations (aside from *All Variations*), the success rate remains stable, suggesting that the model is not overfitting to the training set variations even when the dataset has more samples biased towards those perturbations.

D. How much does each sensory modality contribute to model robustness?

In an effort to investigate the significance of each of the modalities in our system—vision (wristview cameras), touch (force-torque), and proprioception—we conduct an ablation study with models that have one or more input modalities missing. We evaluate each model when trained on a dataset with no variations (i.e. no augmentations) and a dataset with 6 augmentations per demonstration on a training variation set of *Grasp Pose*, *Peg/Hole Shape*, and *Object Body Shape* to analyze how each modality combination reacts when task variations are introduced during training. Figure 7 shows reported % success rate change from the canonical environment success rate averaged over 6 seeds for each variation.
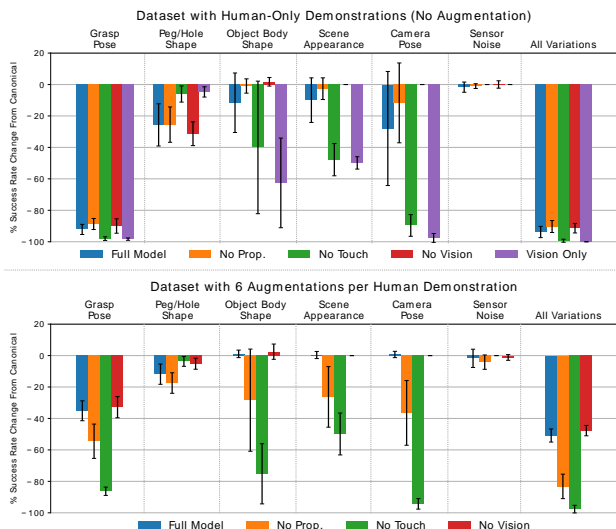

Fig. 7: *% success rate changes on each task variation for models with different modality input combinations trained on no task variations (top) or a subset of task variations (bottom). The* Vision Only *model is omitted from the bottom plot due to training instability. Error bars represent one standard deviation from the mean. The removal of force-torque input sees the largest % success rate drop for many of the task variations out of each of the individual modalities, while the removal of vision has little impact on % success rate change compared to the full model.*

Takeaways: We observe that out of the individual modalities, No Touch has the highest success rate drop for many of the

variations (*Peg/Hole Shape* and *Sensor Noise* being the only exceptions) for both human-only and augmented demonstrations. On the other hand, No Vision has comparable (or sometimes even improved) % success rate changes to Full Model, suggesting its reduced significance in our overall framework compared to the other modalities. Since our task begins immediately in a contact state that is maintained throughout a majority of the task's duration, it follows that force-torque provides the most valuable information about the task state. Proprioception may also give important task state information (as evidenced in the % success rate change for the No Prop. model), as the position of the two end-effectors relative to each other is highly correlated to the position of the peg and hole relative to each other, which is essential knowledge in completing the insertion task. Thus, visual observations provide the least relevant information for our task while still being susceptible to many of the task variations. However, visual input may still be essential in task contexts outside of the one studied here, especially in situations with little to no force-torque feedback (such as aligning the peg and hole objects to be in the same orientation before contact as was studied in our previous work [24]).

VII. CONCLUSION

Towards the goal of evaluating robustness of multisensory policies in contact-rich scenarios such as peg-in-hole object assembly, we present a pipeline for data collection, augmentation, policy training, and evaluation for an object assembly task across diverse observation-level task variations. Our experiments reveal that grasp variations pose the greatest challenge for generalization both in simulation and the real world, and incorporating them through data augmentation significantly improves performance on unseen variations.

While we have demonstrated the ability of our system to learn the underlying task, we acknowledge that the behavior cloning setup used is highly susceptible to covariate shift and cannot recover from erroneous actions. We plan to extend our generalization studies using more advanced imitation learning frameworks, such as Diffusion Policy [30] and ACT [31], and compare their performance with our BC-MLP setup. Moreover, the constrained task initialization and action space in our setup highlight the need to explore more complex, longer-horizon tasks with broader action spaces, and to assess how observation-level task variations affect policies in these contexts.

REFERENCES

[1] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 5026–5033.

[2] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer *et al.*, "Perceiver io: A general architecture for structured inputs & outputs," in *International Conference on Learning Representations*, 2022.

[3] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in *2019 International conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 8943–8950.

[4] M. A. Lee, M. Tan, Y. Zhu, and J. Bohg, "Detect, reject, correct: Crossmodal compensation of corrupted sensors," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 909–916.

[5] Z. Wu, W. Lian, V. Unhelkar, M. Tomizuka, and S. Schaal, "Learning dense rewards for contact-rich manipulation tasks," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 6214–6221.

[6] O. Spector and D. Di Castro, "Insertionnet-a scalable solution for insertion," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5509–5516, 2021.

[7] O. Spector, V. Tchuiev, and D. Di Castro, "Insertionnet 2.0: Minimal contact multi-step insertion using multimodal multiview sensory input," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 6330–6336.

[8] Y. Chen, M. Van der Merwe, A. Sipos, and N. Fazeli, "Visuo-tactile transformers for manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 2026–2040.

[9] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[10] C. Kohler, A. S. Srikanth, E. Arora, and R. Platt, "Symmetric models for visual force policy learning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 3101–3107.

[11] E. Xing, A. Gupta, S. Powers, and V. Dean, "Kitchenshift: Evaluating zero-shot generalization of imitation-based policy learning under domain shifts," in *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021. [Online]. Available: https://openreview.net/forum?id=DdglKo8hBq0

[12] A. Xie, L. Lee, T. Xiao, and C. Finn, "Decomposing the generalization gap in imitation learning for visual robotic manipulation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 3153–3160.

[13] W. Pumacay, I. Singh, J. Duan, R. Krishna, J. Thomason, and D. Fox, "The colosseum: A benchmark for evaluating generalization for robotic manipulation," *arXiv preprint arXiv:2402.08191*, 2024.

[14] J. Gao, A. Xie, T. Xiao, C. Finn, and D. Sadigh, "Efficient data collection for robotic manipulation via compositional generalization," *arXiv preprint arXiv:2403.05110*, 2024.

[15] A. Mandlekar, S. Nasiriany, B. Wen, I. Akinola, Y. Narang, L. Fan, Y. Zhu, and D. Fox, "Mimicgen: A data generation system for scalable robot learning using human demonstrations," in *Conference on Robot Learning*. PMLR, 2023, pp. 1820–1864.

[16] M. Jia, D. Wang, G. Su, D. Klee, X. Zhu, R. Walters, and R. Platt, "Seil: simulation-augmented equivariant imitation learning," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 1845–1851.

[17] L. Ankile, A. Simeonov, I. Shenfeld, and P. Agrawal, "Juicer: Data-efficient imitation learning for robotic assembly," *arXiv preprint arXiv:2404.03729*, 2024.

[18] L. Perez, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.

[19] I. Kostrikov, D. Yarats, and R. Fergus, "Image augmentation is all you need: Regularizing deep reinforcement learning from pixels," *arXiv preprint arXiv:2004.13649*, 2020.

[20] X. Zhu, D. Wang, O. Biza, G. Su, R. Walters, and R. Platt, "Sample efficient grasp learning using equivariant models," in *Robotics: Science and Systems*, 2022.

[21] Z. Chen, S. Kiami, A. Gupta, and V. Kumar, "Genaug: Retargeting behaviors to unseen situations via generative augmentation," *arXiv preprint arXiv:2302.06671*, 2023.

[22] L. Y. Chen, C. Xu, K. Dharmarajan, M. Z. Irshad, R. Cheng, K. Keutzer, M. Tomizuka, Q. Vuong, and K. Goldberg, "Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning," in *Conference on Robot Learning (CoRL)*, Munich, Germany, 2024.

[23] G. Tang, S. Rajkumar, Y. Zhou, H. R. Walke, S. Levine, and K. Fang, "Kalie: Fine-tuning vision-language models for open-world manipulation without robot data," *arXiv preprint arXiv:2409.14066*, 2024.

[24] C. Ku, C. Winge, R. Diaz, W. Yuan, and K. Desingh, "Evaluating robustness of visual representations for object assembly task requiring spatio-geometrical reasoning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 831–837.

[25] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín, "What matters in learning from offline human demonstrations for robot manipulation," in *Conference on Robot Learning*. PMLR, 2022, pp. 1678–1690.

[26] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General perception with iterative attention," in *International conference on machine learning*. PMLR, 2021, pp. 4651–4664.

[27] Y. Zhu, J. Wong, A. Mandlekar, R. Martín-Martín, A. Joshi, S. Nasiriany, and Y. Zhu, "robosuite: A modular simulation framework and benchmark for robot learning," *arXiv preprint arXiv:2009.12293*, 2020.

[28] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2017, pp. 23–30.

[29] B. Mehta, M. Diaz, F. Golemo, C. J. Pal, and L. Paull, "Active domain randomization," in *Conference on Robot Learning*. PMLR, 2020, pp. 1162–1176.

[30] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, 2024.

[31] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.