

HITL-D: Human In The Loop Diffusion Assisted Shared Control

Riley Zilka* and Sergey Khlynovskiy* and Allie Wang* and Martin Jagersand*

Abstract—Autonomous manipulation systems have achieved remarkable capabilities, yet the integration of human expertise with diffusion-based policies in shared control remains relatively unexplored. In this paper, we propose **Human-In-The-Loop Diffusion (HITL-D)**, a shared control framework that enhances user performance in multi-step, insertion, and fine manipulation tasks. HITL-D leverages a novel combination of diffusion-based policies and human control to provide autonomous end effector orientation updates conditioned on a scene point cloud and the Cartesian position of the end effector. This approach reduces the number of joystick control axes required, thereby lowering mental workload. In a multi-task user study with 12 participants, HITL-D reduced average task completion times by 40%, decreased perceived workload by 37%, and improved Likert-scale ratings for independence, intuitiveness, and confidence compared to traditional teleoperation methods. These results demonstrate that HITL-D effectively integrates human expertise with autonomous assistance, improving both objective and subjective aspects of teleoperation.

I. INTRODUCTION

Wheelchair-mounted robotic arms (WMRAs) hold significant promise in enhancing independence and quality of life for individuals with motor impairments [1], [2]. These systems extend the functional capabilities of powered wheelchairs by enabling users to perform activities of daily living (ADLs) such as grasping objects, opening doors, or manipulating devices [3]. However, despite decades of research and technological advancement, WMRAs remain underutilized in practice due to barriers such as cost and accessibility [4]. A major barrier is the difficulty of control without training. Operating a high-degree-of-freedom robotic arm through limited interfaces, such as a standard 2 or 3 degree of freedom wheelchair joystick, imposes substantial mental workload on the user [5]. In the United States alone, more than 9 million people live with a self-care disability [6]. Therefore, developing new systems to reduce mental workload and effort during operation has the potential to improve the quality of life for millions of people.

Shared control is a popular field aiming to reduce workload, improve the ease of control, and minimize degrees of control input. It aims to solve tasks that could be done individually by both the robot and a human, but in such a way that the autonomous and manual inputs from both parties are blended. This reduces the effort by decreasing the necessary dimensions of control and mental effort associated with traditional methods such as Cartesian control [5], [7], [8].

*Department of Computing Science, University of Alberta, Edmonton Alberta, Canada, T6G 2E8. {rzilka, khlynovs, luo3, mj7}@ualberta.ca

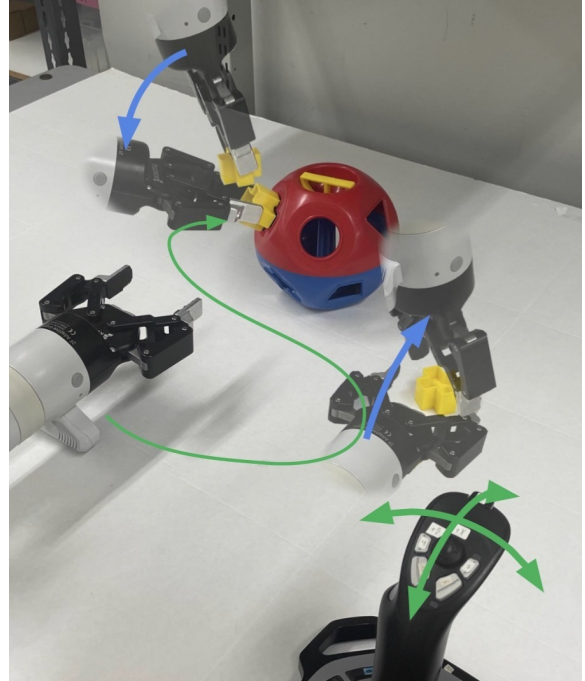


Fig. 1. Our method translates the end effector (green arrow) with 3 joystick-controlled axes, while our shared-control diffusion integration continuously adjusts orientation (blue) based on demonstrations.

Shared control has evolved over several decades, starting with early supervisory frameworks in teleoperation [9], extending to arbitration strategies for assistive WMRAs prioritizing human input in the control and removing complete autonomy [10], and maturing into probabilistic and intent-aware models [11], [12]. More recent approaches increasingly leverage learning-based methods for personalization and high-dimensional control [13]. Learning-based methods are particularly justified by their ability to adapt to user preferences and environment changes while providing autonomous feedback to the system [14]. However, restricting user input, finding efficient mappings, and providing autonomous control can sometimes result in a perceived loss of authority over control [12].

Motivated by the goal of keeping the user in authority at all times while still providing useful autonomous assistance and reducing workload, we present **Human-In-The-Loop Diffusion (HITL-D)**. HITL-D is a simple and effective shared control framework that leverages expert human knowledge and task context, while combining them with the capacity of diffusion policies to represent diverse and complex robot behaviors in a flexible and generalizable format [15], [16].

Similar to the control paradigm in [17], [18], [19] HITL-D works by splitting the two modes of Cartesian control, the current control standard in WMRA: leaving the translation mode (x, y, and z) in the users control while the diffusion policy dictates orientation (yaw, pitch and roll).

In summary, our shared control framework has the following contributions:

- 1) We propose a novel combination of human input and the benefits of diffusion policies for our shared control framework, HITL-D.
- 2) We show that HITL-D is at least as effective as traditional teleoperation methods.
- 3) We show that HITL-D maintains human autonomy and preserves user perceptions of independence.
- 4) We conduct a user study providing evidence that, beyond improved success rates, HITL-D reduces perceived workload compared with traditional teleoperation.

II. RELATED WORK

A. Shared Control

Shared control has been explored in many applications from automotive, medical and assistive robotics [20]. Providing autonomous assistance to an individual during a task while leveraging their intelligent reasoning ability creates a highly effective control paradigm. Many common applications of machine learning models in shared control include path planning [21], [22], intent detection [12], [23], [24] and obstacle avoidance [22], [25]. Beyond solving task constraints, some shared control methods aim to reduce the degrees of freedom the human must control, eliminating cumbersome and sometimes frustrating mode switches [26]. [26] addresses this problem by learning a state conditioned linear mapping through multiple demonstrations to control a 7 DoF robot arm with a 2 DoF joystick controller.

This work does not address motion constraints, including intent detection or object avoidance. There is also no attempt to reduce the robots' degrees of freedom through a reduction mapping. Instead, the user controls the robot's position at all times while a diffusion policy reactively predicts the desired orientation for the task. This approach focuses on maximizing user independence while using shared autonomy to complete tasks faster and with reduced workload.

Similar approaches have addressed orientation control while leaving Cartesian position to the human operator [17], [18]. In [17], control authority is swapped between the human and the autonomous system, whereas our method keeps human authority at all times, with a diffusion policy providing assistance only. [18] combines position and autonomous orientation using guided trajectories from keypoint poses whereas our method preserves the natural mapping of user inputs across all axes. A recent work [19] explores the same orientation–translation split but leverages foundation models and a grasp planner for zero-shot shared control. In contrast, we train a generative policy conditioned on the operator's translational input and a point cloud, producing

orientation corrections that emerge from learned motion distributions without explicit semantic reasoning.

B. Behavior Cloning and Grasp Pose Estimation

Behavior cloning (BC) methods [27], [28], [29] learn policies directly from demonstrations and have been widely applied to robotic manipulation. However, standard BC typically learns a unimodal mapping from observations to actions ($o \rightarrow a$). In shared control settings where behavior is often multi-modal, a single user input could validly lead to multiple distinct manipulation strategies; common BC models often average these modes and suffer from multi-valued function discontinuities [30], resulting in physically inconsistent or ineffective trajectories.

Although grasp pose estimation [31], [32] successfully predicts stable end-state orientations, these static goals often conflict with the user's immediate control priors. Even when providing analytically stable poses, such rigid assistance can diminish the user's sense of agency by overriding their intended manipulation strategy [33].

C. Diffusion Models for Robotics

Diffusion models are an increasingly popular class of generative model in imitation learning and work by treating policy generation as an iterative denoising process [34]. Unlike traditional BC, which often predicts a mean trajectory from its learned unimodal distribution, diffusion policies sample from the joint distribution of entire trajectories, $p(\tau)$. This global perspective provides two critical advantages for shared control: task multi-modality [35], [15] and the ability to seamlessly condition on high-dimensional data [36], [37]. There have been recent advancements conditioning on 3D data in diffusion [35], [38] demonstrating the advantages of including higher dimensional spatial information when conditioning. [35] further demonstrated that these models exhibit superior generalization in high-dimensional spaces compared to standard BC.

While a simple BC model lacks the multi-modal context to fill holes in a trajectory while maintaining temporal coherence, the diffusion process iteratively refines the motion sequence to be consistent with both the learned demonstration experience and the user's real-time guidance. This ensures the system provides orientation assistance that is fluid and goal-oriented. By integrating human input directly into the control loop, we transform the model from an autonomous agent into a collaborative tool that reduces operator workload without sacrificing the expressive capacity of the underlying generative model.

III. METHODS

Human In The Loop Diffusion (HITL-D) is a purely reactive shared control approach where a diffusion policy conditions on sparsely sampled point clouds and end effector positions to predict an end effector orientation as an assisting action. This combination of human input and autonomous output allows the user to maintain control while also benefiting from non-intrusive autonomous assistance.

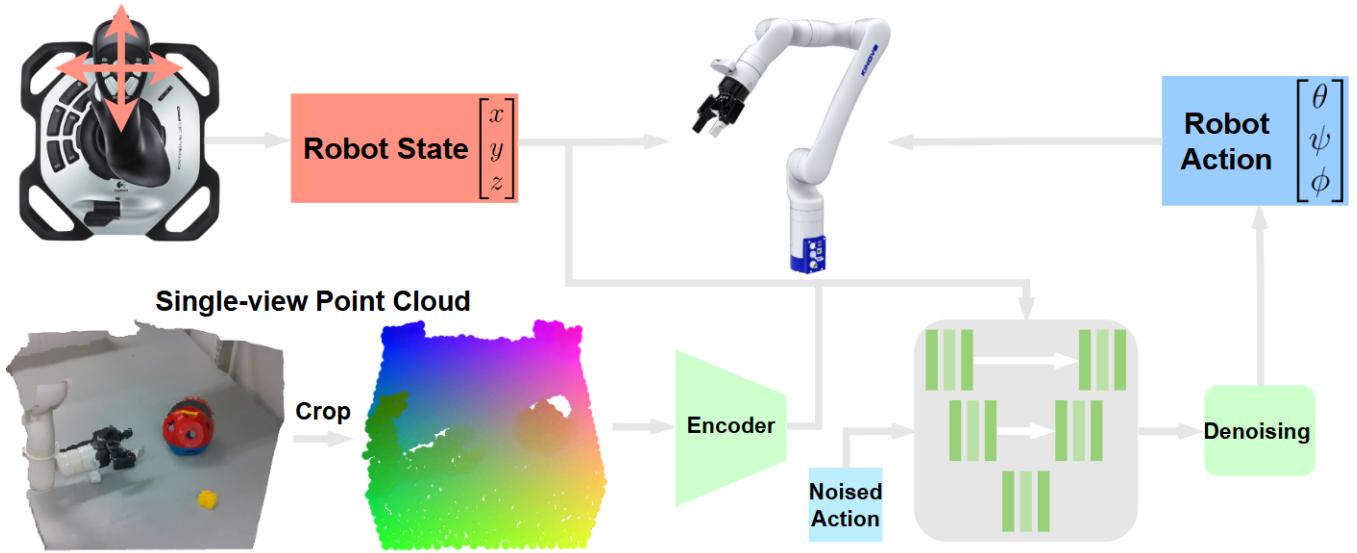


Fig. 2. Overview of HITL-D methodology. During training, expert demonstrations are collected from a teleoperation interface, conditioning our robot action on a cropped and processed point cloud and robot state. During evaluation, a user will manipulate the robot by sending Cartesian position updates to the end effector. These updates run through our control loop along with point clouds to predict end effector orientations.

Orientation was chosen as the action based on two considerations: safety and workload. The user always maintains primary control of the robot as the autonomous predictions cannot change the position of the end effector. Delegating orientation to the autonomous model simplifies system use, allowing the robot to perform fine, task-specific adjustments while the human focuses on coarse positional control. This is especially beneficial for individuals with physical impairments who struggle with fine manipulation.

In HITL-D, our diffusion policy maps observations $o \in O$ to an assisting action $a \in A$ while the user remains in primary control of the robot. Composing our system, there are 3 key parts: (a) **Autonomous Policy Training**, built upon a comparably small number of demonstrations and the perception-decision framework in [35], the action prediction is absolute end effector orientation in radians; (b) **Reactive Diffusion Policy**, adjusting certain hyper-parameters ensures the policy reacts intuitively and preserves user control in real time; (c) **Human-Diffusion Interaction**, an assistive system is developed that reacts to the users motion using techniques that enable rapid, intuitive inference, allowing the system to react predictably to user input. Figure 2 illustrates the overall methodology and control loop.

A. Autonomous Policy Training

The autonomous component of HITL-D is a diffusion policy trained to map multi-modal observations into end effector orientation commands.

Each training step consists of three primary input components:

- **Visual.** HITL-D conditions on a colored point cloud to convey 3D spatial information, improves generalization and reduces demonstration requirements [35]. The point

clouds are cropped with a hand picked bounding box based on the location of the camera relative to the scene. After cropping, farthest point sampling [39] is performed on the point cloud reducing the number of points down to 2048. This number balances inference speed and scene representation quality. The down sampled point clouds are subsequently encoded into a 128 dimension vector for simplified and faster training using the same 3 layer MLP in [35].

- **Robot State.** HITL-D also conditions on a robot state vector containing the 3 dimensional x-y-z position of the end effector relative to the Robot’s base frame.
- **Robot Action.** HITL-D predicts a 3 dimensional end effector orientation in radians.

Our policy is designed as a conditional denoising diffusion model [34], [15] following the same design choices and algorithm as [35] made for their DP3 model.

a) *Action generation:* Similar to DP3, our action is conditioned on 3D visual features v and robot state vector s . Starting from a sample drawn from a Gaussian distribution, the model iteratively denoises this sample until convergence on a single action a . Let a^K denote the initial noise, the denoising network ϵ_θ performs K iterations to produce a noise-free action a^0 according to

$$a^{k-1} = \alpha_k(a^k - \gamma_k \epsilon_\theta(a^k, k, v, s)) + \sigma_k \mathcal{N}(0, I), \quad (1)$$

where $\mathcal{N}(0, I)$ is standard Gaussian noise, and $\alpha_k, \gamma_k, \sigma_k$ are functions of the timestep k defined by the noise scheduler. This iterative procedure corresponds to the *reverse diffusion process* [34].

b) *Training objective:* The network ϵ_θ is trained to predict the noise added to a data point a_0 from the dataset

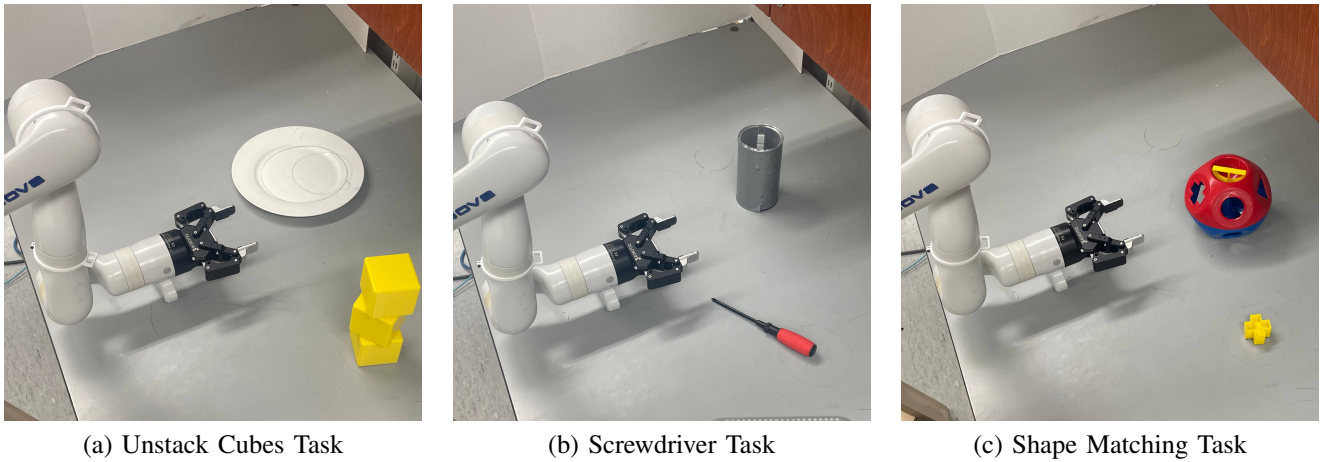


Fig. 3. *Left*: The Unstack Cubes task consists of 3 unaligned cubes and a plate, the goal is to grab each cube on their faces and place them onto the plate such that they are not stacked on the plate. *Middle*: The Screwdriver Task consists of a screwdriver and a gray hollow cylinder, the goal of the task is to grab the screwdriver put it away inside the cup with the handle up. *Right*: The Shape Matching task consists of a cross shape and needs to be carefully aligned with the corresponding hole in the shape matching container.

A. For a given timestep k , the training loss is defined as

$$\mathcal{L} = \text{MSE}\left(\epsilon_k, \epsilon_\theta(\bar{\alpha}_k a^0 + \bar{\beta}_k \epsilon^k, k, v, s)\right), \quad (2)$$

where $\bar{\alpha}_k$ and $\bar{\beta}_k$ are determined by the noise schedule [34].

c) Implementation details: HITL-D adopts a convolutional network-based diffusion policy [15], using DDIM [40] as the noise scheduler. The policy is trained on 3000 epochs with a demonstration length of 256 for all of our tasks, a horizon of 1 and a previous observation size of 1. Hyperparameter explanation is given in III-B.

B. Reactive Diffusion Policy

In pursuit of a real-time, reactive control system, we adjusted the horizon to be 1 action prediction instead of the more commonly used, 8 or 16 [35], [15]. This prioritizes reactivity over guidance. We scrap all future predictions and just orient the end effector in response to the current Cartesian position and point cloud. Previous observation size was reduced to 1 because human motions can be unpredictable, making older observations less informative. We recorded demonstrations with sequence lengths of 256 to ensure the model observes fine-grained temporal transitions, allowing it to predict orientation corrections for a range of typical motions seen in the demonstrations.

One significant positive result of our reactive control design is that the safety is guaranteed and determined by the human’s control over the robot’s Cartesian position; in response, the robot may rotate its joints but it is restricted from autonomously moving to unexpected positions in the workspace, preventing risk of damage to itself or surrounding persons.

C. Human-Autonomy Interaction

The interaction between human control and orientation assistance via our diffusion policy is designed to be straightforward and intuitive. Since the policy predicts a desired orientation, velocity commands are generated based on the

difference between the prediction and current orientations. This provides rapid corrections when large discrepancies exist, while naturally slowing movements as finer adjustments are required. The gain for this simple proportional controller is 0.05. All motion is capped by a velocity limit to maintain safety of operation. This provides a smooth and responsive change for the users while not being quick enough to feel unnatural.

IV. EXPERIMENTS

The most important metric for evaluating the viability of our system is performance with real human users compared to other telemanipulation methods. These experiments were approved by by University of Alberta Research Ethics and Managements (Pro00054665) and signed consent was obtained from the 12 non-disabled participants (ages 19-29, 3 female). Experiments followed a single-blind format with control method randomized across trials. Each participant used each control method for each of the three tasks.

- **Cartesian Control:** This method provides the user with 2 modes, one controls the 3 dimensions of Cartesian space (x, y, z) and the other mode controls end effector orientation (yaw, pitch, roll).
- **Point and Go Control:** This method [41] remaps the coordinate frames of the end effector creating a ”point and go” behavior, it was shown that Point and Go outperformed Cartesian control in most tasks while also being better than the learned SCL method [26].
- **HITL-D:** Our unique shared control method combining diffusion assisted orientation with human translational control.

Each participant was given a brief description of the controls and whether or not it had mode switching, but no context on what the controls did in terms of manipulating the robot. The users had 60 seconds to test and familiarize with each system before attempting to complete 3 trials for each task using each system. The three tasks used in our

TABLE I
QUANTITATIVE RESULTS¹ FROM THE USER STUDY COMPARING CONTROL SYSTEMS FOR THREE TASKS.

Experiment	Metric	Completion Time (s)			NASA-TLX Workload (/100)			Resets (#/Trial)			Success Rate (%)		
		HITL-D	PnG	Cart.	HITL-D	PnG	Cart.	HITL-D	PnG	Cart.	HITL-D	PnG	Cart.
Unstack		95.5±7.1	110.4±8.0	124.2±9.9	32.5±4.8	42.1±5.9	43.8±3.8	0.11±0.09	0.06±0.04	0.28±0.09	100.0	91.7	86.1
Screwdriver		36.5±5.8	88.1±9.5	72.9±9.3	25.5±3.4	53.6±5.2	42.1±6.4	0.22±0.11	0.22±0.09	0.17±0.08	100.0	94.4	97.2
Shape Match		56.7±6.7	126.1±7.0	96.9±7.7	32.0±6.0	57.6±6.1	52.5±7.2	0.33±0.10	0.44±0.09	0.25±0.10	97.2	77.8	97.2
Overall		62.9±19.1	108.2±21.1	98.0±17.4	30.0±14.5	51.1±16.4	46.1±13.8	0.22±0.25	0.24±0.16	0.23±0.24	99.1	88.0	93.5

¹ Results are reported with standard error. Bold = Shared significantly better than Point and Go; Underline = Shared significantly better than Cartesian.

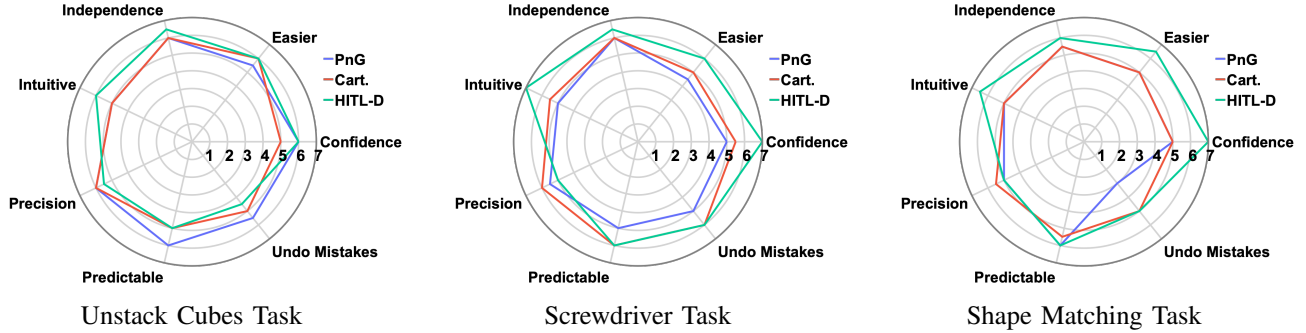


Fig. 4. Radar plots for Likert scale survey measuring individual ratings for different parts of the system. (1-least favorable, 7-most favorable)

study (Fig. 3) were selected to capture different types of ADLs [3] and levels of orientation demand that average WMRA users may face during day to day activities. The unstacking task primarily tested translational control, while the screwdriver and shape matching tasks required frequent orientation adjustments. This design provided insight into whether the benefits of our shared control method extended to both orientation-intensive and orientation-light tasks.

HITL-D’s diffusion policy was trained by first performing a single demonstration with an Xbox 360 controller, with Cartesian position mapped to the left stick and orientation mapped to the right stick with various functions spread across the buttons. This allowed simultaneous control over all degrees of freedom of our Kinova Gen3 Lite robot arm while doing the demonstrations. The advantage of this control setup is that our demonstrations feel smoother and more natural as there is no need to pause and switch between the translation and orientation modes. All demonstrations were performed by an expert researcher to ensure high-quality trajectories. We evaluated and trained our policy on a GeForce RTX 3090Ti.

Notably, our framework requires only one demonstration per task for action prediction, whereas autonomous baselines like DP3 [35] and Diffusion Policy [15] utilize 40 to 284. This reduced data requirement stems from our focus on shared-control rather than fully autonomous end-to-end learning. While we do not claim broad generalization or robustness from a single sample, consistent with broader research in robotic diffusion, we expect that increasing the demonstration count would further enhance policy capability and adaptability.

A. Study Protocol and Metrics

Participants were asked to perform three tasks. For each task, they used three separate control methods across 3 trials. The robot was always in the same starting position and each scene setup was replicated for each user in the study. At the end of each task-method pair, the users were asked to fill out 2 surveys.

1) **Likert Survey:** A 7 point survey measuring the following metrics;

- *Confidence:* how confident are you in using the teleoperation mapping?
- *Easier:* how easy was completing the task with this method?
- *Independence:* how independently do you feel you can control the robot?
- *Intuitive:* how intuitive was the system?
- *Precision:* how precise were the commands in the system?
- *Predictable:* how predictable was the robots behavior?
- *Undo:* how easy was it to undo mistakes with this system?

2) **NASA Task Load Index (TLX):** [42] developed a system to measure subjective workload across different tasks. In this study, it is used to enable meaningful comparisons between methods.

Trial videos, task completion time, successes and failures, mode switches, completion times and the number of resets were collected for this study. However, mode switches were not collected for our HITL-D method as there is only 1 mode. A reset for a task was required when either the cubes

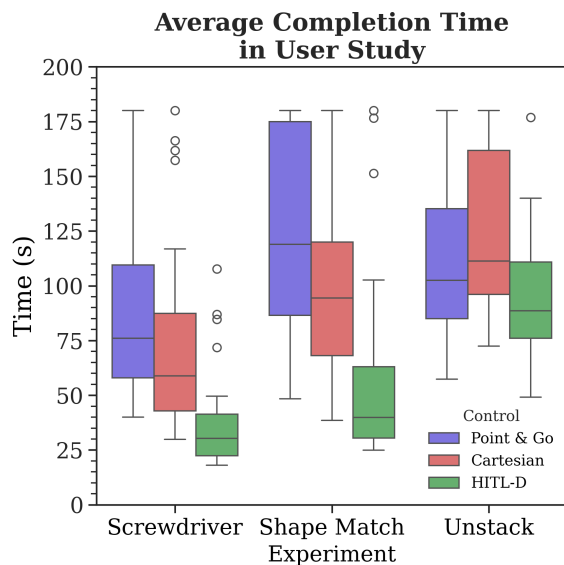


Fig. 5. Box plot for completion times showcasing clearer differences in means, deviations and outliers

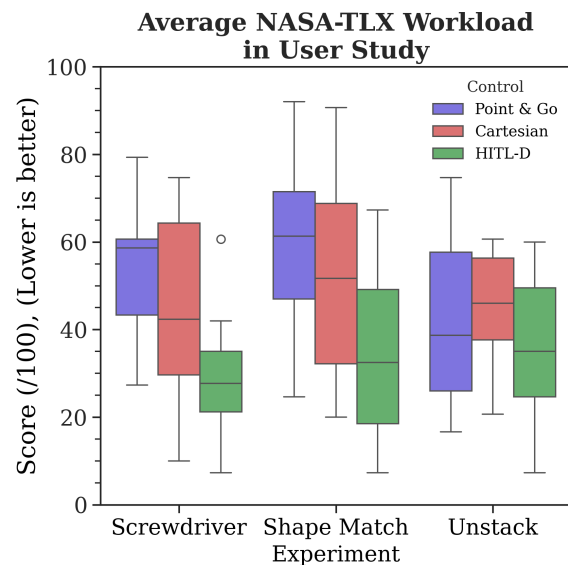


Fig. 6. Box plot for NASA-TLX Workload ratings showcasing clearer differences in means, deviations and outliers

were knocked over or the screwdriver was not dropped in the cylinder and became unreachable. A user could also request a reset if the their control resulted in a robot configuration they could not understand or recover from.

Comparisons of quantitative metrics were achieved with two-way repeated-measures ANOVA. Binary success rates were evaluated with Cochran’s Q test. For discrete data from the Likert assessments, Wilcoxon signed-rank tests were conducted. The threshold for significance was set at 0.05.

V. RESULTS

Table I shows completion times, NASA-TLX workload scores, number of resets and success rates for each task-method pairing. Figure 4 displays the Likert scale metrics for both robot behavior and perceived competence.

Overall we observed a significant decrease in completion times and perceived workloads across all tasks with HITL-D. Notably, standard deviations were lower for each metric with the exception of average resets per trial, indicating that performance improvements were consistent across participants. Figures 5 and 6 demonstrate the visual significance of HITL-D’s improvements over Cartesian and Point and Go.

Unstack Task: HITL-D showed significant improvements with a time reduction of 13.5% and a NASA-TLX workload improvement of 22.8% in comparison to Point and Go. HITL-D was 23.1% faster and had a 25.8% lower workload than Cartesian. HITL-D maintained competitive in the Likert scores of independence and predictability and while not statistically significant, HITL-D showed trends of being more intuitive to the users. We also observed that HITL-D had a 100% success rate while Point and Go and Cartesian methods did not.

Screwdriver Task: We observe significant improvements of HITL-D over both PnG and Cartesian methods in completion times and NASA-TLX workload scores. HITL-D was

nearly twice as fast as Cartesian, reducing average completion time by 49.9%, the workload score was 39.4% lower. HITL-D was more than twice as fast as PnG improving the average time by 58.6% and more than halving the workload score by with an improvement of 52.4%. HITL-D again maintained a success rate of 100% while Cartesian and PnG were close by at above 94% success rates in all trials. HITL-D also showed higher scores, and while the differences were not statistically significant, nearly all Likert survey metrics, except precision, indicate that users preferred HITL-D for this task.

Shape Match Task: Again, we observe significant improvements of HITL-D over both PnG and Cartesian methods in completion times and workload scores. In comparison with Cartesian, HITL-D reduced average completion time by 41.2%, the workload score was 39.0% lower. HITL-D was more than twice as fast as PnG improving the average time by 55.0% and reducing the workload score by 44.4%. HITL-D had a very high success rate of 97.2% and while the PnG struggled to keep up at 77.8% success rate, Cartesian tied HITL-D at 97.2%. HITL-D improved in most metrics in the Likert scale survey except precision which still matched the other two methods.

VI. DISCUSSION

Unstack Task: The comparatively smaller performance gain of HITL-D over Cartesian and PnG methods in the unstacking task can be explained by the fact that this task does not require large or frequent orientation changes, depending on the user’s approach. We observed that some users bumped or pushed the middle cube to rotate it into a position they could grasp while others spent the time switching modes to get the correct orientation. The statistically significant improvement provided by HITL-D arises from its ability to automatically orient with respect to the cubes correctly,

skipping the bumping step and reorienting toward the plate. We make the claim that our method is at least as effective as traditional teleoperation methods. Supporting this claim, the results from the unstack cubes task demonstrate that in situations where orientation adjustments aren't necessary, our method maintains or improves almost all metrics measured in the Likert survey of our method in comparison to the other two.

Screwdriver Task: Our evaluation indicates that users significantly benefited from the shared control framework in this task. Some users picked up the screwdriver completely perpendicular to the table and others picked it up at quite a shallow orientation before the autonomous component could reach its prediction. Despite having only one demonstration, the model was still capable of assisting the user effectively with variations in user strategy leading to a 100% success rate. HITL-D's autonomous component for orientation proved significantly useful in user workload ratings as well as the empirical evidence in completion times. We noticed that the areas users spent the most time in while using Point and Go was when the forearm joints needed to flip due to the change of control frames, which lead to visible frustration. Users who did not hit this case of joint flipping were not as frustrated and were able to complete the task with competitive times. The Cartesian control method also proved to be complex for users who did not use the mode switching for the Unstack task before this one. They were not as familiar with the orientation mode, causing them to perform worse. Users typically failed at this task or needed resets due to dropping the screwdriver outside of the tube. One interesting set of trials during the study led to a false negative for the HITL-D method, during the Cartesian experiment the user would lightly pick up the screwdriver so that it had enough friction to stay in the grippers but just low enough so that it would rotate within them to be completely vertical. When they were unable to get this technique working with HITL-D they assumed that it was the control method that wasn't doing what they desired and kept attempting this behavior and needing to reset. Figure 5 demonstrates these times as outliers as they were the only inconsistent trials during the experiment. Subsequently this false negative resulted in the only outlier for the workload rating seen in Figure 6. Even with this outlier HITL-D remains significantly well suited for a task like this that necessitates large orientation changes.

Shape Match Task: This task posed the greatest difficulty for users, as achieving a successful insertion required the end effector orientation to be nearly exact in order to align with the shape-matching container. Users commonly struggled with the roll of the end effector for the Cartesian method which caused the cross to be unaligned. Similar to the screwdriver task, users would pick up the cross at a very shallow angle, resulting in them not having enough pitch available in the workspace to orient the end effector so the cross had the right orientation. We also noticed a particular behavior only for the HITL-D method due to its shared control nature. Users inexperienced with the method and task often attempted to pick up the plus shape at a shallow angle.

Upon trying to insert it they found that they were constrained by the end effectors ability to orient correctly with various pick up orientations and would have to reset. Evidence for this behavior is demonstrated in Figure 5 which visualizes the outliers in trial times. Once users recognized that the autonomous model required an additional period to achieve an optimal orientation when picking up the cross, they were able to complete the task more quickly, and no resets were needed after the first trial.

Across All Tasks: Overall, the observations demonstrate the efficacy of our method in time to completion, perceived workload, and on average higher user preferences for our method in the Likert metrics such as intuitiveness, ease of control and independent ability to move the robot. An interesting result we observed, although not statistically significant, was that Point and Go showed trends in higher task completion times and workload ratings than Cartesian in the Screwdriver and Shape Match tasks. We attribute this to the fact that Point and Go is a recent novel control method [41] and that while it may require more time to learn, it shows promising results for experienced users. Most importantly, we proved our claim that our method does not reduce human autonomy: users experienced increased autonomy and independence compared to that of Point and Go and Cartesian methods.

VII. CONCLUSION

In this work, a novel combination of diffusion and human input for robot teleoperation is proposed, where a user controls a robot via a joystick and sends Cartesian velocity commands and a diffusion policy reacts to the Cartesian pose to provide an end effector orientation to assist the user in a reactionary setting. Through a user study, users maintained a sense of independent control while experiencing a significant reduction in mental workload using our framework. We observed statistically significant improvements in task completion, highlighting the efficiency and usability of our approach, especially with reduced-DoF controllers. These findings emphasize the benefits of shared control over purely human or autonomous control, demonstrating the unique advantages of combining diffusion-based fine manipulation capabilities with human task expertise. Future work includes conducting a user study on impaired individuals to validate the usability and workload reduction for our intended demographic. Additionally, evaluating a wider variety of ADLs would give better insight into the applicability of this method. Further, we acknowledge the limitation of a hand cropped point cloud and want to explore additional ways to improve invariance to object, camera and robot arm variations. Another direction is incorporating visual recognition models to better localize objects in the scene, which could enable benefits like dynamic cropping and new forms of conditional data.

REFERENCES

- [1] B. Driessen, H. Evers, and J. v Woerden, "Manus—a wheelchair-mounted rehabilitation robot," *Proceedings of the Institution of Me-*

- chanical Engineers, Part H: Journal of engineering in medicine, vol. 215, no. 3, pp. 285–290, 2001.
- [2] V. Maheu, P. S. Archambault, J. Frappier, and F. Routhier, “Evaluation of the jaco robotic arm: Clinico-economic study for powered wheelchair users with upper-extremity disabilities,” in *2011 IEEE international conference on rehabilitation robotics*. IEEE, 2011, pp. 1–5.
 - [3] L. Petrich, J. Jin, M. Dehghan, and M. Jagersand, “A quantitative analysis of activities of daily living: Insights into improving functional independence with assistive robotics,” in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 6999–7006.
 - [4] J. Bourassa, J. Faieta, J. Bouffard, and F. Routhier, “Wheelchair-mounted robotic arms: a survey of occupational therapists’ practices and perspectives,” *Disability and Rehabilitation: Assistive Technology*, vol. 18, no. 8, pp. 1421–1430, 2023.
 - [5] L. V. Herlant, R. M. Holladay, and S. S. Srinivasa, “Assistive teleoperation of robot arms via automatic time-optimal mode switching,” in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 35–42.
 - [6] “Centers for disease control and prevention.” [Online]. Available: <https://www.cdc.gov/dhds/about/index.html>
 - [7] H. A. Tijsma, F. Liefhebber, and J. L. Herder, “Evaluation of new user interface features for the manus robot arm,” in *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005*. IEEE, 2005, pp. 258–263.
 - [8] A. Campeau-Lecours, U. Côté-Allard, D.-S. Vu, F. Routhier, B. Gosselin, and C. Gosselin, “Intuitive adaptive orientation control for enhanced human–robot interaction,” *IEEE Transactions on robotics*, vol. 35, no. 2, pp. 509–520, 2018.
 - [9] T. B. Sheridan, *Telerobotics, automation, and human supervisory control*. MIT press, 1992.
 - [10] T. Carlson and Y. Demiris, “Human-wheelchair collaboration through prediction of intention and adaptive assistance,” in *2008 IEEE international conference on robotics and automation*. IEEE, 2008, pp. 3926–3931.
 - [11] A. D. Dragan and S. S. Srinivasa, “A policy-blending formalism for shared control,” *The International Journal of Robotics Research*, vol. 32, no. 7, pp. 790–805, 2013.
 - [12] S. Javdani, S. S. Srinivasa, and J. A. Bagnell, “Shared autonomy via hindsight optimization,” *Robotics science and systems: online proceedings*, vol. 2015, pp. 10–15 607, 2015.
 - [13] K. Muelling, A. Venkatraman, J.-S. Valois, J. E. Downey, J. Weiss, S. Javdani, M. Hebert, A. B. Schwartz, J. L. Collinger, and J. A. Bagnell, “Autonomy infused teleoperation with application to brain computer interface controlled manipulation,” *Autonomous Robots*, vol. 41, no. 6, pp. 1401–1422, 2017.
 - [14] A. Broad, T. Murphey, and B. Argall, “Learning models for shared control of human-machine systems with unknown dynamics,” *arXiv preprint arXiv:1808.08268*, 2018.
 - [15] 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*, p. 02783649241273668, 2023.
 - [16] Z. Wang, J. J. Hunt, and M. Zhou, “Diffusion policies as an expressive policy class for offline reinforcement learning,” *arXiv preprint arXiv:2208.06193*, 2022.
 - [17] Z. Jin and P. R. Pagilla, “Collaborative operation of robotic manipulators with human intent prediction and shared control,” in *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, 2020, pp. 1–6.
 - [18] D. Sun and Q. Liao, “Asymmetric bilateral telerobotic system with shared autonomy control,” *IEEE Transactions on Control Systems Technology*, vol. 29, no. 5, pp. 1863–1876, 2021.
 - [19] Z. Fu, P. Song, Y. Hu, and R. Detry, “Tasc: Task-aware shared control for teleoperated manipulation,” *arXiv preprint arXiv:2509.10416*, 2025.
 - [20] D. A. Abbink, T. Carlson, M. Mulder, J. C. F. de Winter, F. Amiravan, T. L. Gibo, and E. R. Boer, “A topology of shared control systems—finding common ground in diversity,” *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 5, pp. 509–525, 2018.
 - [21] P. Bevilacqua, M. Frego, E. Bertolazzi, D. Fontanelli, L. Palopoli, and F. Biral, “Path planning maximising human comfort for assistive robots,” in *2016 IEEE conference on control applications (CCA)*. IEEE, 2016, pp. 1421–1427.
 - [22] G. Quere, A. Hagenbruber, M. Iskandar, S. Bustamante, D. Leidner, F. Stulp, and J. Vogel, “Shared control templates for assistive robotics,” in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 1956–1962.
 - [23] M. S. Erden and T. Tomiyama, “Human-intent detection and physically interactive control of a robot without force sensors,” *IEEE Transactions on Robotics*, vol. 26, no. 2, pp. 370–382, 2010.
 - [24] S. Jain and B. Argall, “Probabilistic human intent recognition for shared autonomy in assistive robotics,” *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 1, pp. 1–23, 2019.
 - [25] J. Storms, K. Chen, and D. Tilbury, “A shared control method for obstacle avoidance with mobile robots and its interaction with communication delay,” *The International Journal of Robotics Research*, vol. 36, no. 5-7, pp. 820–839, 2017.
 - [26] M. Przystupa, K. Johnstonbaugh, Z. Zhang, L. Petrich, M. Dehghan, F. Haghverd, and M. Jagersand, “Learning state conditioned linear mappings for low-dimensional control of robotic manipulators,” *arXiv preprint arXiv:2410.21441*, 2024.
 - [27] F. Torabi, G. Warnell, and P. Stone, “Behavioral cloning from observation,” *arXiv preprint arXiv:1805.01954*, 2018.
 - [28] 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,” *arXiv preprint arXiv:2108.03298*, 2021.
 - [29] F. Codevilla, E. Santana, A. M. López, and A. Gaidon, “Exploring the limitations of behavior cloning for autonomous driving,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9329–9338.
 - [30] P. Florence, C. Lynch, A. Zeng, O. A. Ramirez, A. Wahid, L. Downs, A. Wong, J. Lee, I. Mordatch, and J. Tompson, “Implicit behavioral cloning,” in *Conference on robot learning*. PMLR, 2022, pp. 158–168.
 - [31] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, “Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics,” *arXiv preprint arXiv:1703.09312*, 2017.
 - [32] A. Ten Pas, M. Gualtieri, K. Saenko, and R. Platt, “Grasp pose detection in point clouds,” *The International Journal of Robotics Research*, vol. 36, no. 13-14, pp. 1455–1473, 2017.
 - [33] M. A. Collier, R. Narayan, and H. Admoni, “The sense of agency in assistive robotics using shared autonomy,” in *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2025, pp. 880–888.
 - [34] J. Ho, A. Jain, and P. Abbeel, “Denosing diffusion probabilistic models,” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
 - [35] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, “3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations,” *arXiv preprint arXiv:2403.03954*, 2024.
 - [36] Y. Wang, Z. Liu, L. Yang, and P. S. Yu, “Conditional denosing diffusion for sequential recommendation,” in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2024, pp. 156–169.
 - [37] Z. Zhan, D. Chen, J.-P. Mei, Z. Zhao, J. Chen, C. Chen, S. Lyu, and C. Wang, “Conditional image synthesis with diffusion models: A survey,” *arXiv preprint arXiv:2409.19365*, 2024.
 - [38] T.-W. Ke, N. Gkanatsios, and K. Fragkiadaki, “3d diffuser actor: Policy diffusion with 3d scene representations,” *arXiv preprint arXiv:2402.10885*, 2024.
 - [39] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.
 - [40] J. Song, C. Meng, and S. Ermon, “Denosing diffusion implicit models,” *arXiv preprint arXiv:2010.02502*, 2020.
 - [41] A. Wang, C. Jiang, M. Przystupa, J. Valentine, and M. Jagersand, “Point and go: Intuitive reference frame reallocation in mode switching for assistive robotics,” in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 1–7.
 - [42] S. G. Hart and L. E. Staveland, “Development of nasa-tlx (task load index): Results of empirical and theoretical research,” in *Advances in psychology*. Elsevier, 1988, vol. 52, pp. 139–183.