Understanding Control Frames in Multi-Camera Robot Telemanipulation

Pragathi Praveena, Luis Molina, Yeping Wang, Emmanuel Senft, Bilge Mutlu, and Michael Gleicher
Department of Computer Sciences, University of Wisconsin–Madison, Madison, Wisconsin, USA
{pragathi, lamolina, yeping, esenft, bilge, gleicher}@cs.wisc.edu

***Abstract*—In telemanipulation, showing the user multiple views of the remote environment can offer many benefits, although such different views can also create a problem for control. Systems must either choose a single fixed control frame, aligned with at most one of the views or switch between view-aligned control frames, enabling view-aligned control at the expense of switching costs. In this paper, we explore the trade-off between these options. We study the feasibility, benefits, and drawbacks of switching the user's control frame to align with the actively used view during telemanipulation. We additionally explore the effectiveness of explicit and implicit methods for switching control frames. Our results show that switching between multiple view-specific control frames offers significant performance gains compared to a fixed control frame. We also find personal preferences for explicit or implicit switching based on how participants planned their movements. Our findings offer concrete design guidelines for future multi-camera interfaces.**

***Index Terms*—telemanipulation, camera, awareness, control frame, multiple views, operator interfaces**

I. Introduction

Robot telemanipulation extends human capacity by allowing users to explore and physically affect a remote environment. During telemanipulation, visibility of the remote workspace is necessary for users to complete tasks safely and successfully. To provide visibility, telerobotics interfaces commonly use video streams from one or more remote cameras [1]. Multiple viewpoints of the remote workspace can improve remote perception by providing cues about scale, depth, and spatial relations between elements in the remote environment and alternate sources of information in the presence of occlusions. However, controlling a robot using multiple viewpoints can be a challenge because the various views may not align with the coordinate frames of the user's input device or the robot in the remote workspace, resulting in loss of spatial orientation and poor teleoperation performance. Prior work considering a single viewpoint has shown that aligning the *control frame*, the frame of reference in which the user provides motion commands, with the viewpoint can improve performance because of the consistency in the direction of the user's movement and the consequent movement of the robot on the video display [2–6]. However, extending this idea to multi-camera interfaces requires answering questions regarding which view the control frame should be aligned with and whether, and how, the control frame should be changed when the user switches between views.

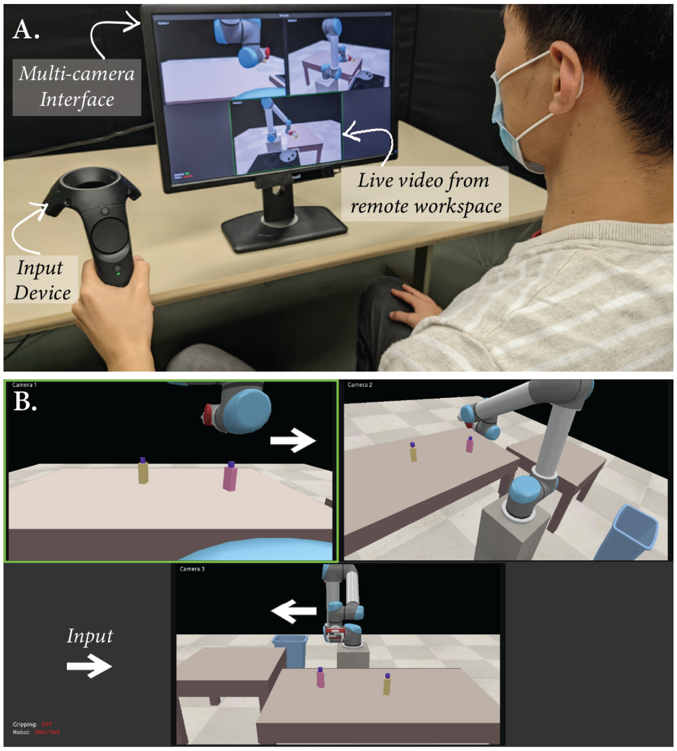

Fig. 1. In this paper, we investigate the effects of using multiple view-specific control frames in a multi-camera interface on task performance and user experience during robot telemanipulation. **A.** In our user study, participants controlled a robot arm in simulation to complete tasks that followed a home-care scenario using live video feeds from three static cameras placed in the remote environment. **B.** In a multi-camera interface, viewpoints with different frames of reference can negatively impact performance. Here, the control frame is aligned with the top-left view (green border). User input to the right results in robot movement that is rightward in the top-left view, but leftward in the bottom view. This mismatch makes it challenging to control the robot using the bottom view and a control frame aligned with the top-left view.

Previous multi-camera interfaces have primarily used a *single* control frame aligned with the robot's base or one of the views [5, 7, 8]. This approach requires the user to perform mental transformations when the control frame and the view frame do not align. We posit that using *multiple* control frames aligned with the different views will reduce the mental transforms required. However, given that a single control frame can be used at any given time, users must switch control frames based on which view frame they wish to use, which may also introduce additional control burden. Automated approaches, particularly an attention-based (using eye-tracking) approach, to control-

This research was supported in part by NSF award 1830242 and NASA University Led Initiative (ULI) grant 80NSSC19M0124. Code and videos related to this work are available at https://pages.graphics.cs.wisc.edu/MultiView.

frame switching might alleviate this burden. In this paper, we investigate the advantages and limitations of using *multiple* control frames in multi-camera interfaces and examine two approaches, *manual* and *attention-based*, to changing control frames when the user switches views.

We conducted a user study to investigate the effects of using multiple view-specific control frames on telemanipulation performance. We compared a baseline interface with a single control frame against two interfaces with multiple view-specific control frames. In our baseline interface, the control frame was aligned with the static frame of the robot's base joint, offering a one-to-one mapping from the user's space to the robot's space that remains consistent across various viewpoints. However, using viewpoints that are not aligned with the robot requires additional mental transformations for control, increasing cognitive load and reducing task performance [4, 9]. The other two interfaces offered access to three control frames corresponding to three camera views. These view-specific control frames allow for visual consistency between the user's input and the robot's movements in each view. However, switching between control frames requires the user to reorient from one frame to another. For the interfaces with multiple control frames, we designed two methods for real-time selection of the view that the control frame is aligned with: (1) *explicit* view selection by the user manually pressing a button; and (2) *implicit* (or automatic) view selection based on the view that the user is attending to as inferred from eye-gaze data. Our results show that using multiple view-specific control frames offers significant performance gains compared to a single control frame. Additionally, user preferences for explicit or implicit view selection depend on how they plan their movements.

The central contribution of our work is the demonstration of the feasibility and benefits of using multiple view-specific control frames in multi-camera telemanipulation interfaces, which has implications for the design of multi-view interfaces.

II. Related Works

A. Control Frames

The control frame, or the frame of reference in which the user provides motion commands, is an important design decision for teleoperation interfaces that can affect task performance and user experience. In comparisons of three widely used control frames, aligned with the frame of the robot's base (*robot frame*), of the viewpoint (*view frame*), or of the target object (*task frame*), no one frame has emerged as the "best" [9]. These control frames have trade-offs that make them more or less suitable for different applications. Because task frames are bespoke to the specific objects to be manipulated [10], in this work, we aligned control frames with the robot and with the view frames to maximize the generalizability of our findings.

Based on prior work, we discuss the trade-off between control frames aligned with the robot and view frames. A control frame aligned with the robot's base is easy to define with respect to the robot and allows spatial correspondence between the user's and the robot's workspaces (e.g., [11, 12]). However, if the user's viewpoint is not aligned with the robot, additional mental transformations are required for the user to predict the result of their inputs [4, 13]. Hence, a control frame aligned with the viewpoint (e.g., [4, 6]) can be beneficial because the robot, as seen from that viewpoint, moves in the same direction as the user's motion command and requires no mental transformations. While a viewpoint-aligned control frame results in a lower cognitive load, it can be less efficient (requiring more motion commands) due to its indirect spatial relationship to the robot's workspace [9]. Note that this discussion of trade-offs in prior work is limited to using a *single* control frame aligned with the robot or view frame.

B. Multi-Camera Telerobotics Interfaces

Studies of factors that affect human performance in teleoperation systems have noted that multiple viewpoints improve remote perception, but integrating information across viewpoints with different frames can have a negative impact [1]. Nielsen et al. [14] addressed this issue by creating a single integrated view by spatially combining information from sources with different frames. Similarly, other works have displayed one video feed within another in a manner that is spatially consistent with the locations of the cameras in the remote environment [15, 16]. These solutions enable the user to control the robot in a single frame of the reference of the integrated view. While prior work has effectively combined two views, scaling this approach to an arbitrary number of disparate views, which can be necessary for complex telemanipulation, is challenging and not always feasible. Another approach to combining multiple views is to display a spatial map of the remote scene on a head-mounted display [17–19], providing the user with access to different vantage points by moving through a 3D scene. This approach can offer a more immersive and intuitive experience than displaying multiple video feeds. However, due to the technical and computational challenges of high-quality real-time 3D reconstruction, directly streaming video feeds is still the prevalent approach in large, cluttered, and dynamic settings, such as in construction [8] or search and rescue [20].

Existing multi-camera interfaces (e.g., [7–9, 15, 20]) commonly use a single frame of reference aligned with the base of the robot or one of the views. Keyes et al. [21] highlighted the limitations of a single frame of reference. Their robot was equipped with a front-facing and a rear-facing camera. The interface allowed an Automatic Direction Reversal (ADR) mode that reversed the commands along the front-back axis when using the rear-facing camera. This mode enabled users to control the robot with fewer collisions. In the absence of this mode, the performance of the two-camera system was similar to an interface with only the front-facing camera. This result suggests that users may not effectively utilize an additional viewpoint if their control frame is misaligned with the view frame. Similarly, a study of the use of multiple camera views during laparoscopic surgery found that while multiple views improved task performance, the view that was most misaligned with the control frame provided the least benefit [22].

Based on this literature that underscores the limitations of a single frame of reference in multi-camera interfaces, we propose using *multiple* view-aligned control frames. As discussed in §II-A, a view-aligned control frame has been shown to be beneficial in single-camera interfaces. We posit that those benefits will extend to multi-camera interfaces and outweigh any costs associated with switching from one control frame to another. Keyes et al. [21] consider control frame switches in their interface; however, the control frame change in their work was limited to one translational axis (front-back). In contrast, our work considers three control frames with differences in multiple translational and rotational axes to address a knowledge gap in the existing literature on the costs of switching between disparate view-specific control frames.

C. View Selection

Our proposed solution requires users to switch between control frames based on which view frame they wish to use. To enable this feature, similar to work by Keyes et al. [21], we implemented *explicit* view selection where the robot movement depends on the view that the user manually selects with a button. However, given the high cognitive demands of telemanipulation [1], we posit that there will be benefits to automating the view selection. Because users are visually attending to the views, we consider the use of the *direction of gaze* of the operator as an automated and implicit way of aligning control frames with the currently attended view. To further develop this idea, we draw from HCI literature on attention-based interfaces [23, 24] and gaze-based selection [25, 26]. We implemented the simplest design for *implicit* view selection, where the robot movement depends on the view the user is currently looking at as determined by the measured direction of gaze. Jacob et al. [27] note that people often move their eyes for reasons that may or may not be task-related. Given the lack of prior work on gaze-based switching of control frames, it is unclear if the advantages offered by automation outweigh unintended control frame switches due to movements that are not task-related. Our work aims to address this knowledge gap.

III. Design and Prototype

In this section, we discuss the design of our *implicit* and *explicit* view selection approaches and implementation details of our prototype system for our study with human subjects.

A. Design of Active View Selection

We define *active view* as the view with which the control frame is aligned. We identify two costs for switching the active view. The first cost is *a loss in momentum*. Switching from one control frame to another may require the user to give different motion commands for the same robot action. For example, imagine two control frames, *C1* and *C2*, that have their left-right axis flipped with respect to each other (refer Figure 1.B). For simplicity, let *C1* be aligned with the robot's base frame. To move the robot to its left, the user would need to move the controller left in *C1* and right in *C2*, as the left-right axis is flipped in the second control frame. Suppose that the user is trying to reach a target to the left of the robot. The user starts by using *C1* and moves the controller to the left to move the robot leftward. When the user switches to *C2*, the user needs to now move the controller to the right to maintain the robot's leftward trajectory. This reorientation of the controller command leads to a *loss in momentum*. The second cost is the additional *control burden* of providing an appropriately-timed input for active view selection.

The trade-off between *explicit* and *implicit* active view selection approaches is shaped by the trade-off between the control burden and loss in momentum. Explicit view selection requires the user to deliberately select a view for control (the user may look at a view without selecting it), whereas during implicit view selection, looking toward a view automatically aligns the control frame with that view. Implicit view selection allows fast and easy selection but results in a loss of momentum. Explicit view selection requires deliberate selection but allows the user to prepare for the change in the control frame.

B. Prototype Details

1) Teleoperation Control Interface: Our system used *RelaxedIK*, a mimicry-control teleoperation approach proposed by Rakita et al. [11, 28] that allows novice users to effectively control a robot arm using full six-DoF arm-space control. An HTC VIVE motion controller served as the input device to capture user commands mapped to the robot's movement. Prior work [11, 29] suggests that free-form controllers such as a VR controller are significantly better for telemanipulation than alternatives such as a touch interface or keyboard and mouse. Our interface allowed *clutching*, the ability to disengage control and move the input device independent of the robot (similar to lifting a finger and repositioning it on a trackpad). We implemented this capability using the grip button on the VIVE controller. In addition, we configured the controller's trigger to toggle the opening and closing of the robot gripper.

2) Multi-Camera Interface: We implemented a multi-camera interface using OpenGL and ImGui that allows a variable number of video feeds to be displayed in various layouts (such as Picture-in-Picture or Grid layout). For this work, we used a grid layout to display live video feeds from three cameras (see Figure 1.B). Both layout and number of cameras are key design decisions for multi-camera interfaces, and our choices represent only one possible combination. We chose a grid layout because it does not emphasize one camera and concurrently shows all feeds. We chose three cameras because a pilot study suggested that three cameras were adequate for our telemanipulation tasks but presented sufficient complexity to necessitate frequent switching of control frames. The *active view* or the view with which the control frame was aligned was highlighted with a green border (see top-left view in Figure 1.B).

3) Active View Selection: For view selection, we implemented two approaches. For explicit view selection, we spatially mapped regions on the trackpad of the VIVE controller to corresponding views: *Trackpad Left* → *Camera 1*, *Trackpad Right* → *Camera 2*, *Trackpad Down* → *Camera 3*. This mapping allowed the user to switch their active view with one click on the trackpad. For implicit view selection, we used the Tobii Pro X2-60 screen-based eye tracker along with the Tobii Pro

SDK to determine the view toward which the user was looking and selected that view as the active view. If the user's gaze fixation did not fall within the boundaries of one of the views (for example, when the user looks away, looks at the blank space between views, or the eye tracker does not detect the eyes), then the last used view continued to be the active view.

4) Simulated Workspace: We chose to use a simulated workspace for our evaluation to have easy access to arbitrary viewpoints and maintain the robot's safety during shifts in control frames. We simulated the workspace on the CoppeliaSim [30] platform with the ODE physics engine and created separate scenes for each task in the user study. Users controlled a six-DoF Universal Robots UR5 robot arm equipped with a parallel gripper. We implemented all dynamic task objects using native shapes in CoppeliaSim to allow fast simulation by the physics engine. Our implementation tried to maintain realistic physics that the user would encounter in an actual teleoperation scenario. For example, if the robot was holding an object in its gripper and collided with the table, the object could drop from the gripper due to the collision force. For real-time performance in CoppeliaSim, the graphical rendering of the workspace was simple and lacked realistic shadows and textures. Because multiple views are the root cause of the control frame issues we are addressing, we do not anticipate the graphics quality to impact our findings significantly. However, we note that recently developed platforms such as NVIDIA Isaac Sim can offer more realistic simulations.

5) Viewpoint Selection: We added three exocentric vision sensors to the simulated workspace corresponding to each task. We chose viewpoints to provide three requirements: (1) adequate coverage of the workspace, (2) detail and context of the objects to be manipulated, and (3) depth perception. Across all tasks, *Camera 1* was placed above the robot's base, aligned with the static frame of the base joint. This view was selected as the active view in the condition where the user operated the robot in a single fixed control frame. The other two viewpoints were different for each task to fulfill the requirements listed above. Video feeds from the simulated cameras are shown in real-time on the multi-camera interface described in §III-B2.

6) Control Frames: We used a right-handed Cartesian coordinate system to represent all frames of reference. We obtained the camera's orientation matrix in the simulated environment relative to the base of the robot. In real life, this matrix can be obtained from camera calibration using a checkerboard pattern. Based on prior work [6, 31], we constructed the view frame matrix to have camera-right and world-up axes. Thus, the view frame consists of three orthogonal vectors: (1) the camera's left-right axis; (2) the direction of gravity for the up-down axis; and (3) for the front-back axis, a normalized vector perpendicular to both aforementioned axes. When a view is selected as the active view, the user's control input is transformed to match that frame.

C. System Architecture

We implemented our teleoperation system on a single computer running Linux OS using Robot Operating System (ROS). The user provides inputs to the system through a VIVE controller and Tobii eye tracker. The outputs from the system were robot joint states from *RelaxedIK* that were used to operate the robot arm in a simulator, three views of the remote workspace displayed on a 24-inch video monitor, and the selection of an active view corresponding to the control frame.

IV. User Study

We conducted a user study to assess how control frame choices affect teleoperation performance and user experience.

A. Hypotheses

> **H1:** In a multi-camera telemanipulation interface, enabling the user to switch between several view-specific control frames (either explicitly or implicitly) will result in better task performance and user experience than a single fixed control frame.

§II-B discussed issues with control and viewpoint usage in existing interfaces that use a single fixed control frame. We expect the benefits of using a view-aligned control frame in single-camera interfaces to extend to multi-camera interfaces and outweigh the costs of switching from control frames.

> **H2:** In a multi-camera telemanipulation interface with several view-specific control frames, implicit (or automatic) view selection will result in better task performance and user experience than explicit view selection.

Due to the high cognitive demands of telemanipulation, we expect automating view selection to result in better outcomes.

B. Experimental Design

To test our hypotheses, we designed a 3 × 1 within-participants experiment in which participants completed telemanipulation tasks using (1) *fixed frame*, (2) *adaptive frame (explicit)*, and (3) *adaptive frame (implicit)* in a fully counterbalanced order. In *fixed frame*, the control frame, which was aligned with the static frame of the robot's base joint and *Camera 1*'s view, remained fixed during the tasks. In *adaptive frame (explicit)*, the control frame was aligned with the active view frame that the user selected by pressing the appropriate region on the controller's trackpad. In *adaptive frame (implicit)*, the control frame was automatically aligned with the active view frame based on the direction of the user's gaze.

C. Tasks

We designed three pick-and-place tasks based on home care scenarios: *cleanup*, *meal-serve*, and *meal-prep*. *Cleanup* and *meal-serve* had to be completed in three minutes, and *meal-prep* in four. The tasks were ordered to increase in difficulty.

The *cleanup* task involved picking up two bottles and dropping them in the trash bin placed at a distance. This task was easy in terms of manipulation but challenging in terms of tracking broad motions across the different views.

The *meal-serve* task involved picking up a sandwich and a bottle from a shelf and placing them on a tray for service. Any collisions with the shelf were converted into audible warning beeps to discourage collisions. This task was challenging due to the need to maintain spatial awareness to avoid collisions while reaching an ergonomically inconvenient location and precisely orienting the gripper to grasp the objects.

In the *meal-prep* task, participants picked up two small jars with meal ingredients (two cubes) and emptied them into a pan. This task was challenging because of the rotational dexterity required to empty the content of the jars.

All tasks asked participants to interact with two objects in a pre-specified order to maintain consistency across participants. Participants heard an auditory confirmation for each subtask they completed. For the *cleanup* and *meal-serve* tasks, each object to be manipulated was considered a subtask resulting in two subtasks each. Each cube in the jar was considered a subtask for the *meal-prep* task, resulting in four subtasks.

We designed a *training task* with a time limit of 7.5 minutes that included all the objects that the participants would interact with in the main tasks but placed them at different locations.

D. Procedure

The procedure was administered under a protocol reviewed and approved by the Institutional Review Board (IRB) of University of Wisconsin–Madison. Following informed consent, the experimenter introduced participants to the idea of remotely controlling a robot in someone's kitchen to finish up their chores and guided them through an interactive training session to familiarize themselves with teleoperation. Then, participants completed training for each condition, where the experimenter explained the control frame choices for that condition and provided details about each task. Participants were informed that the tasks were challenging and that they should get through as much of the tasks as possible in the time allotted and aim for performing the tasks well. If participants had a catastrophic failure, such as dropping the object on the ground where it is inaccessible by the robot, they were asked to move on to the next object. After completing three tasks in the current condition, participants filled out a questionnaire regarding their experience. After all three conditions, participants completed a demographics survey and engaged in a semi-structured interview about their overall experience. During the interview, participants were asked to articulate their planning and execution strategies in each condition and pick one condition that they preferred overall and reason about it. These responses were recorded and transcribed by an experimenter.

E. Measures

1) Performance measures: Participant task performance reflected the average task score of the three tasks. For each task, we calculated the normalized binary success over its subtasks. Thus, the possible scores for *cleanup* and *meal-serve* were {0, 0.5 and 1}, and for *meal-prep* were {0, 0.25, 0.5, 0.75, 1}.

2) Subjective measures: We developed a 10-item questionnaire (Table I) and derived three scales that measured *perceived ease of use* (items 3, 6, 10; Cronbach's α = 0.89), *perceived predictability* (items 1, 4, 8; Cronbach's α = 0.83), and *perceived spatial orientation* (items 2, 5, 7, 9; item 5 reversed; Cronbach's α = 0.77). Participants responses were captured using a seven-point rating scale (1–7; 1 = "Strongly Disagree," 7 = "Strongly Agree"). We took an unweighted average of ratings of the items on the NASA Task Load Index (TLX) to measure perceived workload [32].

TABLE I
QUESTIONNAIRE ITEMS ADMINISTERED AFTER EACH CONDITION

#	Item
1	The robot's motion was not surprising.
2	I was aware of what was happening in the remote environment.
3	The control method made it easy to accomplish the task.
4	The robot responded to my motion inputs in a predictable way.
5	I felt disoriented while completing the tasks.
6	I felt confident controlling the robot.
7	I always knew how to get to my desired location in the remote environment.
8	The robot consistently moved in a way that I expected.
9	I felt confident about where things were located in the remote environment.
10	I could accurately control the robot.

3) Behavioral measures: We recorded screen coordinates derived from eye gaze data using Tobii Pro X2-60 and all inputs provided by the user through the HTC VIVE controller.

F. Participants

We recruited 24 participants (17 male, 7 female) from a university campus between the ages of 18 and 39 (M = 21.7, SD = 5.29). Participants reported some familiarity with robots (M = 3.33, SD = 1.46, measured on a seven-point scale). Two participants reported an interaction with a robot in prior robotics research studies. The study took 90 minutes, and all participants received $20 USD as compensation.

V. Results and Discussion

Our data analysis provides partial support for our hypotheses, and we discuss these results in §V-A. Interestingly, the data distribution across all the measures showed high inter-subject variability. To better understand these variations between participants, we conducted a *post hoc* data analysis discussed in §V-B. This analysis identified groups of participants who had similar strategies for planning their movement, which helps us better characterize the cost of switching control frames.

A. Results

Our analysis used one-way repeated-measures analysis of variance (ANOVA), considering the control method as the within-participants factor for each of the outcome measures described in §IV-E. If there was a significant effect, we used pairwise t-tests with Bonferroni correction to determine where the differences lied. These results are summarized in Figure 2.

Our results provide partial support for *H1*. Both *adaptive frame* conditions resulted in significantly higher task scores than the *fixed frame* condition. However, the results for user perception effects were mixed. Scores for ease of use, predictability, and spatial orientation were higher in the *adaptive frame (implicit)* condition than the *fixed frame* condition, but no significant differences were found between the *adaptive frame (explicit)* condition and the *fixed frame* condition. For perceived workload, there were no significant differences between the

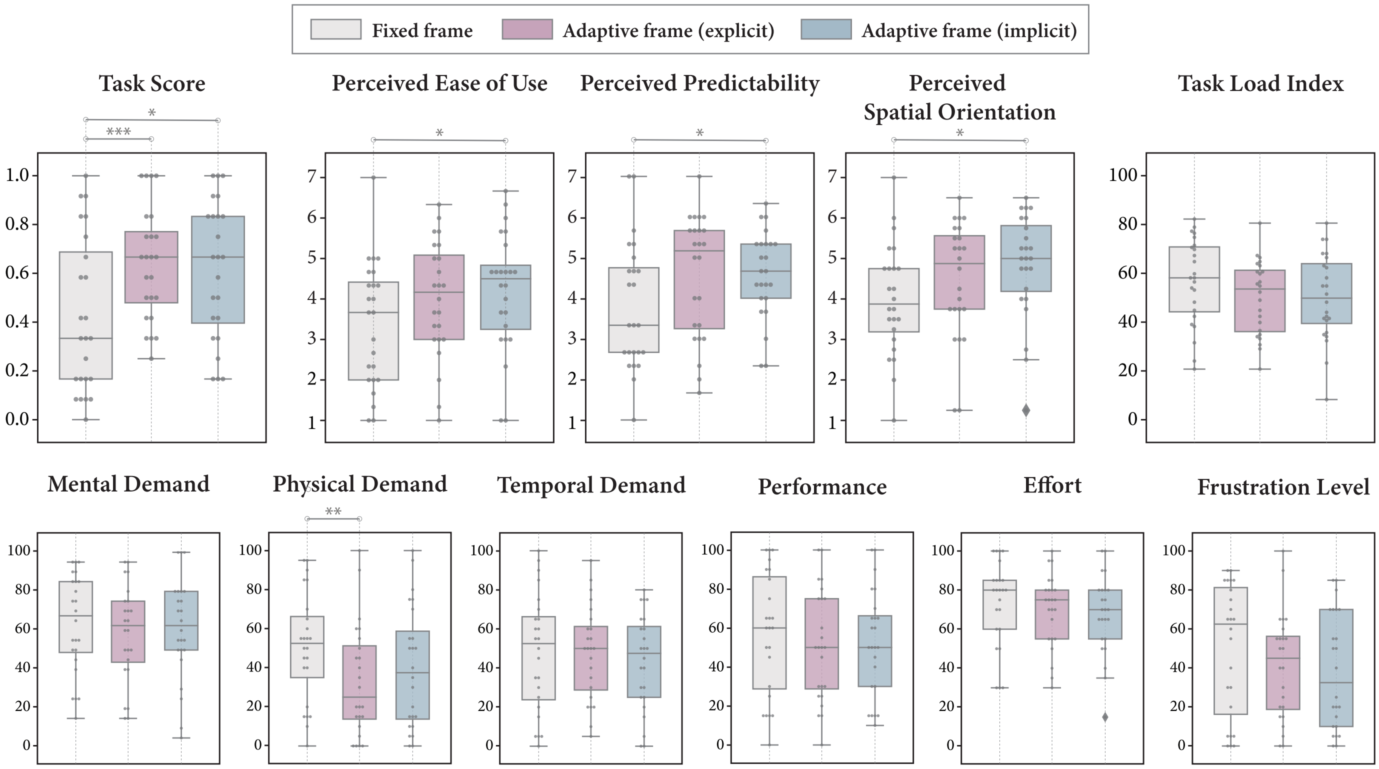

Fig. 2. Tukey boxplots overlaid on data points from performance and user perception measures. Horizontal lines indicate significant pairwise comparisons with Bonferroni correction ($p < 0.05^*$, $p < 0.01^{**}$, $p < 0.001^{***}$).

three conditions for the TLX score. However, among the subscale ratings, perceived physical demand was significantly lower for the *adaptive frame (explicit)* condition than the *fixed frame* condition. Finally, we found no support for *H2*: no significant differences were observed between the two *adaptive frames*.

For all measures, we tested for order effects using repeated-measures analysis of covariance (ANCOVA) considering ordinal position (first, second, third) as the covariate and found no significant effects of the order of the control frames.

B. Post hoc Analysis and Discussion

Motivated by the high inter-subject variability in our dataset, we conducted a multimodal exploratory analysis using additional data collected during the study, such as motion data, eye gaze data, and responses from semi-structured interviews. Our findings, summarized in Table II allow us to characterize control frame choices in multi-camera interfaces better. We discuss the findings in connection with our two hypotheses.

We report participant numbers *(P1–P24)* in our findings, with participants sorted based on the best task score among the three conditions. Thus, *P1* finished the least proportion of tasks in any condition, while *P24* finished the most.

1) Fixed vs. Adaptive Frames: Participants in Group 1 *(G1: P12, P20, P21, P22, P23, P24)* successfully used or preferred the *fixed frame*. These participants employed a distinct strategy that sometimes resulted in more efficient movements than when using *adaptive frames* but likely led to higher physical demand. We discuss the analysis that supports this finding below.

Our observations suggest that not clutching played a vital role in the performance outcomes of successful participants in the *fixed frame* condition. As explained in §III-B1, clutching allows the user to move the controller independent of the robot. We counted instances of clutching across all tasks in the *fixed frame* condition and found a medium negative correlation (Pearson's r = –0.56) with task scores. As seen in Figure 3.A, many successful participants *(P20, P21, P22, P23, P24)* rarely used the clutch. These participants indicated that resuming robot control after clutching required them to rethink the mapping from their workspace to the robot's workspace, which they avoided altogether by not clutching. Clutching enables operators to reorient. In the *fixed frame* condition, reorientation can result in loss of spatial orientation and predictability, particularly in rotational movements. Not clutching led to awkward arm positions, as seen in Figure 3.C, which may explain the increased physical demand scores compared to the *adaptive frame (explicit)* condition.

In the interviews, four participants *(P12, P20, P21, P22)* chose *fixed frame* as their overall preferred condition because they perceived it to be more efficient for task completion. Averaging task times of completed subtasks for the five most successful participants *(P20, P21, P22, P23, P24)* showed that using *adaptive frame (explicit)* and *adaptive frame (implicit)* took 32% and 43% longer, respectively, compared to using *fixed frame*. Participant comments highlight a drawback of using multiple control frames that require frequent reorientation. Because clutching can be utilized for reorientation, we averaged the clutching counts for the five participants. We found that, on average, they used clutching 34, 24, and 5 times in *adaptive frame (explicit)*, *adaptive frame (implicit)*, and *fixed frame*

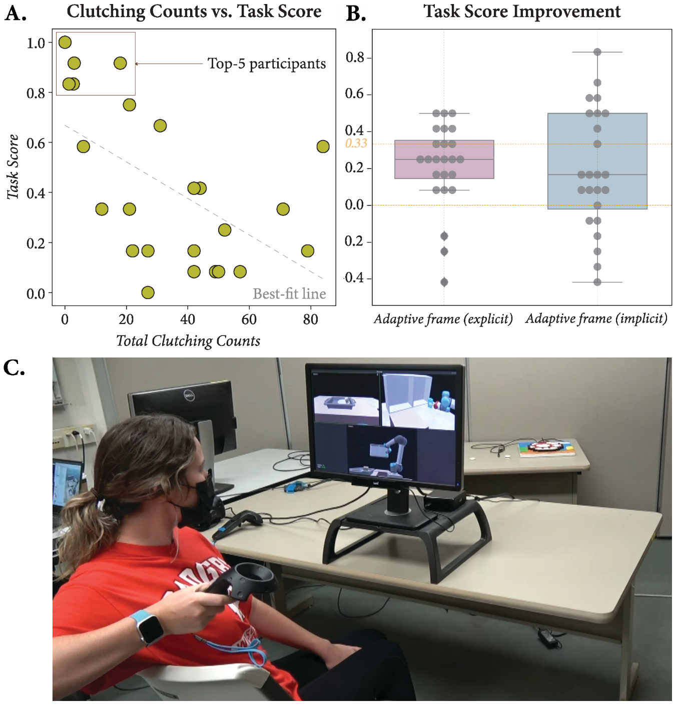

Fig. 3. **A.** Scatterplot of the number of times clutching was utilized across all tasks in the *fixed frame* condition with respect to task scores. We found a medium negative correlation (Pearson's r = –0.56) between the measures. **B.** Tukey boxplots of improvements in task scores in the *explicit* and *implicit* conditions compared to the *fixed frame* condition. Our *post hoc* analysis dives deeper into the strategies employed by participants who had high performance gains (*Task Score Improvement* $>$ 0.33) and saw performance losses (*Task Score Improvement* $<$ 0) compared to the baseline. **C.** A participant completing manipulations in an awkward arm position in the *fixed frame* condition.

conditions, respectively. This frequent reorientation could contribute to the inefficiency of multiple view-specific frames. However, efficiency is not the only objective in telerobotics systems: all five participants completed all tasks using one of the *adaptive frames*, but only one *(P22)* did using *fixed frame*.

In summary, a few participants (*G1* in Table II) could efficiently complete some tasks with a *fixed frame* by employing a physically demanding strategy. However, even these participants could complete more tasks using *adaptive frames*.

2) Implicit vs. Explicit Active View Selection: While there were no significant differences between the *adaptive frame* conditions across measures, we observed some differences in the distribution of the data. A multimodal analysis of our dataset identified participant groups who used similar strategies for planning their movement. Table II summarizes the different groups and strategies. Our findings suggest that users perceive the trade-off between the costs associated with switching control frames differently depending on their strategy.

To compare the *adaptive frame* conditions, we first calculated improvement in task scores in the *explicit* and *implicit* conditions compared to the *fixed frame* condition (Figure 3.B). Consistent with the results presented in §V-A, a majority of participants had performance gains *(Task Score Improvement* $>$ 0) when using either of the *adaptive frame* conditions *(explicit:* 21/24 participants, *implicit:* 17/24 participants).

Further, we examined data of nine participants in each condition *(explicit: P6, P8, P9, P10, P13, P16, P17, P18, P19, implicit: P6, P7, P8, P9, P10, P15, P16, P17, P18)* who had high performance gains compared to the *fixed frame* condition *(Task Score Improvement* $>$ 0.33). Seven out of nine participants *(explicit:* not *P13, P19, implicit:* not *P7, P15)* had high performance gains in both conditions, and six out of these seven participants *(*not *P8)* performed better in the *implicit* condition. Note that performance gains for this group cannot solely be attributed to learning effects since the ordinal position of the *implicit* condition was first for three participants and last for the other three participants. To sum up, a substantial number of participants who benefited highly from using the *adaptive frame* conditions preferred automatic view selection.

Data from the nine participants *(P6, P7, P8, P9, P10, P15, P16, P17, P18)* who had high performance gains in the *implicit* condition indicate two distinct groups of participants based on the strategy they used to complete the tasks. Group 2 *(G2: P10, P15, P16, P17, P18)* was more likely to manually switch their active view when they looked at a new view. Because this function was automatic in the *implicit* condition, this group saw performance gains over the *explicit* condition. This group chose *implicit* as their overall preferred condition. Group 3 *(G3: P6, P7, P8, P9)* often forgot to switch their active view when they were engrossed in the task, which resulted in mistakes such as collisions and dropping objects. Only one participant *(P8)* in this group performed better in the *explicit* condition but had a better subjective experience in the *implicit* condition. This group was equally split in their overall preferences *(explicit: P6, P7, implicit: P8, P9)*. One participant *(P24)* completed all tasks in both *adaptive frame* conditions but preferred automatic camera selection because it allowed for more fluid interaction. Overall, *G2* and *G3* (Table II) benefited from the *implicit* condition in either or both subjective and objective outcomes.

We also examined the data from six participants *(P11, P12, P13, P14, P21, P23)* who saw performance losses in the *implicit* condition compared to the baseline *(Task Score Improvement* $<$ 0). Except for one participant *(P12)*, all other participants completed the most tasks in the *explicit* condition and chose it as their overall preferred condition. To this group of five participants, we added one participant *(P19)* who also completed the most tasks in the *explicit* condition and chose it as their preferred condition but did not have performance losses in the *implicit* condition compared to the baseline. The strategy of this group *(G4: P11, P13, P14, P19, P21, P23)* was to make a minimal number of switches to their active view only when absolutely necessary.

The other group that completed the most tasks in the *explicit* condition and chose it as their overall preferred condition were the participants with the lowest task scores, Group 5 *(G5: P1, P2, P3, P4, P5)*. This group preferred the convenience of manipulating in view-specific control frames. However, they often switched between views (only for visual information, not with the intention of switching control frames), making control too jarring with automatic view selection. Overall, *G4* and *G5* (Table II) were discouraged by the frequent loss in momentum

TABLE II

PARTICIPANT STRATEGIES AND PREFERRED EXPERIMENTAL CONDITIONS FELL INTO FIVE GROUPS

Group	Participants	Strategy	Participant Quote	Favorable Frame
G1	P12, P20, P21, P22, P23, P24	Rarely utilized clutching, benefited from efficient spatial mapping between the user's and robot's workspaces with the single fixed frame	*P20:* I am a pretty spatial person. Once I put myself in the robot's space, it *(fixed frame)* was easy. I couldn't pause *(clutch)* the robot though.	Fixed
G2	P10, P15, P16, P17, P18, P22	Switched active view frequently, benefited from automatic selection	*P15:* With B *(explicit)* I had to stop moving and switch the camera, but C *(implicit)*, I just did it, without stopping, without having to think at all	Adaptive (implicit)
G3	P6, P7, P8, P9, P24	Forgot to manually switch active view when focusing on the manipulation task, benefited from attention-based selection	*P9:* I thought B *(explicit)* was irritating, a distraction almost. Thinking about which button I want to press and then pressing it, was so much work	Adaptive (implicit)
G4	P11, P13, P14, P19, P21, P23	Experienced loss in momentum during gaze-based selection, preferred minimal switching of the active view when absolutely necessary	*P11:* With B *(explicit)*, I could look at a different camera but still move in the direction I originally planned with another camera	Adaptive (explicit)
G5	P1, P2, P3, P4, P5	Experienced loss in momentum during gaze-based selection, preferred having time to think and prepare for switching the active view	*P4:* My brain was so tripped up by how I move my hand and how the arm moved on the screen, and I couldn't process it fast enough in C *(implicit)*	Adaptive (explicit)

in the *implicit* condition and preferred the *explicit* condition to switch the active view when they deemed it a value-addition.

While our *post hoc* analysis identified groups of participants with similar strategies for planning their movement, we have no additional insight into why participants chose these strategies. Analysis of demographic data, which included aggregate scores of participant familiarity with video games, joystick controllers, VR controllers, and 3D modeling software by groups (*G1 – G5*) did not offer meaningful insights.

VI. General Discussion

In this paper, we explore the use of multiple view-specific control frames to improve telemanipulation using multi-camera interfaces, highlighting its feasibility, advantages, and limitations. The current practice of using a single control frame requires the user to integrate information from various views with different frames of reference and maintain a *global* spatial mental model of the robot's workspace. In contrast, switching between view-specific control frames allows the user to work in a *local* frame of reference without maintaining global spatial awareness. However, switching control frames comes at a cost, namely the control burden of selecting the appropriate view to align the control frame with and a loss in momentum from changing the frame of reference of the input commands. Users perceive the trade-off between these costs differently depending on their strategy (refer Table II).

Implications for Design of Future Multi-view Interfaces:

1) Enabling switching between multiple view-aligned control frames allows users to work in a local frame of reference without maintaining global spatial awareness.
2) A view offers visual information as well as a control frame to move through the workspace. Therefore, when picking a suitable viewpoint for a task, it may be worthwhile to optimize not only for the visual objectives but also for a reasonable frame of reference for the task.
3) Both manual and attention-based switching are feasible approaches that can be beneficial in different ways depending on the user. For example, some users experience an additional control burden while switching between frames and benefit from attention-based switching. Other users experience a significant loss of momentum during control and prefer to switch frames manually.
4) While attention-based switching is promising, the simple design we provided needs improvement to benefit a broader range of people. One possible approach is using a two-step selection process such as EyePoint [33], where eye gaze is used to reference a view, and manual input is used to select it as the active view. Other potential approaches are Pinpointing [34] and ReType [35].

Limitations – Our study has some limitations that must be addressed by future work. First, participants were given minimal training to perform the study tasks, and how their strategies may evolve with more training and long-term use must be explored further. Second, our study followed a within-participants study design, which required participants to learn and adapt to different control frames and switching methods in a single 90-minute session. A study with separate sessions for each condition on different days or a between-participants design would provide participants with more time to build expertise on each method. Third, our study involved a six-DoF robot arm as the minimum requirement for unconstrained traversal of a 3D workspace, and participant strategies, including their ability to adapt to different control frames, might differ across robot designs. Because our proposed control frame switches facilitate visual consistency between user input and robot movements in each view, we expect these results to translate to higher DoF robots. Future work can investigate the effects of robot morphology on how teleoperators adapt to different control frames. Fourth, given the knowledge gap on the impact of switching between disparate control frames, we were unsure about the effect of control frame changes on the safety of robot motion and thus utilized a simulated environment. While we expect our findings to translate to real-world scenarios, further research is needed to substantiate this expectation. Finally, our multi-camera interface included three static cameras with bespoke viewpoints. Future work can examine how additional views, different types of views, and dynamic cameras affect the costs for switching control frames.

References

[1] J. Y. Chen, E. C. Haas, and M. J. Barnes, "Human performance issues and user interface design for teleoperated robots," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 6, pp. 1231–1245, 2007.

[2] J. A. Macedo, D. B. Kaber, M. R. Endsley, P. Powanusorn, and S. Myung, "The effect of automated compensation for incongruent axes on teleoperator performance," *Human Factors*, vol. 40, no. 4, pp. 541–553, 1998.

[3] K. Chintamani, A. Cao, R. D. Ellis, and A. K. Pandya, "Improved telemanipulator navigation during display-control misalignments using augmented reality cues," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, no. 1, pp. 29–39, 2009.

[4] B. DeJong, J. Colgate, and M. Peshkin, "Mental transformations in human-robot interaction," in *Mixed Reality and Human-Robot Interaction*. Springer, 2011, pp. 35–51.

[5] M. Draelos, B. Keller, C. Toth, A. Kuo, K. Hauser, and J. Izatt, "Teleoperating robots from arbitrary viewpoints in surgical contexts," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 2549–2555.

[6] D. Rakita, B. Mutlu, and M. Gleicher, "An autonomous dynamic camera method for effective remote teleoperation," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 325–333.

[7] S. Hughes, J. Manojlovich, M. Lewis, and J. Gennari, "Camera control and decoupled motion for teleoperation," in *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483)*, vol. 2. IEEE, 2003, pp. 1339–1344.

[8] R. Sato, M. Kamezaki, S. Niuchi, S. Sugano, and H. Iwata, "Cognitive untunneling multi-view system for teleoperators of heavy machines based on visual momentum and saliency," *Automation in Construction*, vol. 110, p. 103047, 2020.

[9] L. M. Hiatt and R. Simmons, "Coordinate frames in robotic teleoperation," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2006, pp. 1712–1719.

[10] D. H. Ballard, "Task frames in robot manipulation." in *AAAI*, vol. 19, 1984, p. 109.

[11] D. Rakita, B. Mutlu, and M. Gleicher, "A motion retargeting method for effective mimicry-based teleoperation of robot arms," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017, pp. 361–370.

[12] O. Porges, M. Connan, B. Henze, A. Gigli, C. Castellini, and M. A. Roa Garzon, "A wearable, ultralight interface for bimanual teleoperation of a compliant, whole-body-controlled humanoid robot," in *IEEE International Conference on Robotics and Automation ICRA*. IEEE, 2019.

[13] L. Wu, F. Yu, J. Wang, and T. N. Do, "Camera frame misalignment in a teleoperated eye-in-hand robot: Effects and a simple correction method," *arXiv preprint arXiv:2105.08466*, 2021.

[14] C. W. Nielsen, M. A. Goodrich, and R. W. Ricks, "Ecological interfaces for improving mobile robot teleoperation," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 927–941, 2007.

[15] S. H. Seo, D. J. Rea, J. Wiebe, and J. E. Young, "Monocle: interactive detail-in-context using two pan-and-tilt cameras to improve teleoperation effectiveness," in *2017 26th IEEE international symposium on robot and human interactive communication (RO-MAN)*. IEEE, 2017, pp. 962–967.

[16] H. Hedayati, M. Walker, and D. Szafir, "Improving collocated robot teleoperation with augmented reality," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 78–86.

[17] L. Peppoloni, F. Brizzi, C. A. Avizzano, and E. Ruffaldi, "Immersive ros-integrated framework for robot teleoperation," in *2015 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE, 2015, pp. 177–178.

[18] J. I. Lipton, A. J. Fay, and D. Rus, "Baxter's homunculus: Virtual reality spaces for teleoperation in manufacturing," *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 179–186, 2017.

[19] D. Whitney, E. Rosen, D. Ullman, E. Phillips, and S. Tellex, "Ros reality: A virtual reality framework using consumer-grade hardware for ros-enabled robots," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–9.

[20] X. Xiao, J. Dufek, and R. R. Murphy, "Autonomous visual assistance for robot operations using a tethered uav," in *Field and Service Robotics*. Springer, 2021, pp. 15–29.

[21] B. Keyes, R. Casey, H. A. Yanco, B. A. Maxwell, and Y. Georgiev, "Camera placement and multi-camera fusion for remote robot operation," in *Proceedings of the IEEE International Workshop on Safety, Security and Rescue Robotics*. National Institute of Standards and Technology Gaithersburg, MD, 2006, pp. 22–24.

[22] P. R. DeLucia and J. A. Griswold, "Effects of camera arrangement on perceptual-motor performance in minimally invasive surgery." *Journal of Experimental Psychology: Applied*, vol. 17, no. 3, p. 210, 2011.

[23] D. D. Salvucci and J. R. Anderson, "Intelligent gaze-added interfaces," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2000, pp. 273–280.

[24] J. S. Shell, R. Vertegaal, and A. W. Skaburskis, "Eyepliances: attention-seeking devices that respond to visual attention," in *CHI'03 extended abstracts on Human factors in computing systems*, 2003, pp. 770–771.

[25] J. Hild, E. Peinsipp-Byma, M. Voit, and J. Beyerer, "Suggesting gaze-based selection for surveillance applications," in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2019, pp. 1–8.

[26] K. B. Moorthy, M. Kumar, R. Subramanian, and V. Gandhi, "Gazed–gaze-guided cinematic editing of wide-angle monocular video recordings," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–11.

[27] R. Jacob and S. Stellmach, "What you look at is what you get: gaze-based user interfaces," *interactions*, vol. 23, no. 5, pp. 62–65, 2016.

[28] D. Rakita, B. Mutlu, and M. Gleicher, "Relaxedik: Real-time synthesis of accurate and feasible robot arm motion," in *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018.

[29] D. Whitney, E. Rosen, E. Phillips, G. Konidaris, and S. Tellex, "Comparing robot grasping teleoperation across desktop and virtual reality with ros reality," in *Robotics Research*. Springer, 2020, pp. 335–350.

[30] E. Rohmer, S. P. N. Singh, and M. Freese, "Coppeliasim (formerly v-rep): a versatile and scalable robot simulation framework," in *Proc. of The International Conference on Intelligent Robots and Systems (IROS)*, 2013.

[31] L. Wu, "Design and implementation of intuitive human-robot teleoperation interfaces," Ph.D. dissertation, University of South Florida, 2020.

[32] 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.

[33] M. Kumar, A. Paepcke, and T. Winograd, "Eyepoint: practical pointing and selection using gaze and keyboard," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2007, pp. 421–430.

[34] M. Kytö, B. Ens, T. Piumsomboon, G. A. Lee, and M. Billinghurst, "Pinpointing: Precise head-and eye-based target selection for augmented reality," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–14.

[35] S. Sindhwani, C. Lutteroth, and G. Weber, "Retype: Quick text editing with keyboard and gaze," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–13.