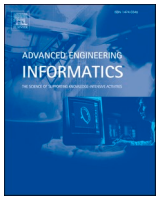
Advanced Engineering Informatics 68 (2025) 103665

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei



Learning viewpoint control from human-initiated transitions for teleoperation in construction

Sungboo Yoon [a], Moonseo Park [a], Changbum R. Ahn [b,*]

[a] Department of Architecture and Architectural Engineering, Seoul National University, Seoul 08826, Republic of Korea
[b] Department of Architecture and Architectural Engineering, Institute of Construction and Environmental Engineering, Seoul National University, Seoul 08826, Republic of Korea

ARTICLE INFO

Keywords:
Imitation learning
Teleoperation
Viewpoint control
Virtual reality
Construction robotics

ABSTRACT

Visual perception is critical for teleoperation in construction, where optimal visibility directly impacts task performance. Hybrid viewpoint control systems enhance the flexibility of visual perception by adaptively coupling or decoupling the viewpoint from robot movements according to situational demands. However, determining the optimal timing for transitions between these perspectives remains a major challenge, as existing autonomous methods are not directly applicable to hybrid control for construction tasks. In this work, we propose a viewpoint control mode prediction model that autonomously manages transitions during teleoperation with hybrid control. Our learning scheme with a transition-guided weighting method leverages sporadic transition commands from human interactions with the teleoperation system as demonstration data for imitation learning. User evaluation in a virtual reality (VR) environment simulating construction welding tasks shows that our model outperforms the baselines, achieving an 11% improvement over the state-of-the-art behavioral cloning (BC) algorithm and a 19% improvement over the state-of-the-art weighted BC algorithm in replicating human transition behaviors. This work contributes novel insights into the design of visual perception systems for teleoperation in construction, enabling reliable, user-aligned viewpoint transitions.

1. Introduction

Teleoperation plays an important role in construction, including reduced human exposure to risks [1–3], data collection for imitation learning [4–6], and on-the-fly adjustments for autonomous systems [7]. Most teleoperation systems rely on single or multiple cameras that are either fixed in position [6] or mounted in an eye-in-hand configuration [8]. These cameras are typically positioned in a task-agnostic manner, without specific consideration for the visibility requirements of particular tasks [9]. Although several perception systems [10–13] include cameras that allow orientational movements, their movements remain constrained by fixed camera setups [14], often suffering from occlusions in cluttered workspaces or when close-up views are required [9]. This limited visual feedback not only hinders precise and efficient remote task execution but also raises safety concerns in construction [15].

To offer more flexibility to remote perception in teleoperation, recent studies have explored dynamic viewpoint systems that allow real-time adjustments to operator views. Dynamic cameras, using external cameras mounted on unmanned aerial vehicles (UAVs) [16,17] or high-degree-of-freedom (DoF) robotic arms [9,14,18–23], typically operate in either robot-coupled or decoupled modes [24]. In the robot-coupled mode, the operator's viewpoint moves along with the robot, providing a coordinated view that aligns closely with the robot's actions. In the robot-decoupled mode, the viewpoint is entirely independent of the robot's movements, allowing greater flexibility in adjusting the viewpoint as needed. In our previous work [25], we introduced a hybrid viewpoint control system that combines both robot-coupled and decoupled behaviors, enabling operators to adapt their viewpoints during teleoperated construction tasks. Construction welding tasks, featuring spatially dispersed weld seams, provide illustrative examples of such applications. When navigating construction sites, operators can employ a robot-decoupled viewpoint to effectively explore the surrounding environment, helping them avoid obstacles and plan efficient lift paths [26]. When performing precise welding operations, operators can switch to a robot-coupled viewpoint to simplify hand-eye coordination and focus on the task at hand [27].

* Corresponding author at: Department of Architecture and Architectural Engineering, Institute of Construction and Environmental Engineering, Seoul National University, Seoul 08826, Republic of Korea.
E-mail addresses: yoonsb24@snu.ac.kr (S. Yoon), mspark@snu.ac.kr (M. Park), cbahn@snu.ac.kr (C.R. Ahn).

https://doi.org/10.1016/j.aei.2025.103665
Received 27 January 2025; Received in revised form 11 July 2025; Accepted 15 July 2025
Available online 21 July 2025

Although hybrid viewpoint control systems offer advantages over individual robot-coupled and decoupled viewpoint control, they introduce new challenges, particularly in managing transitions between the two perspectives. Ideally, transitions should happen at times that are both predictable and minimally disruptive to the operator's workflow. However, defining and automating these transitions is challenging due to their subjective nature. While giving operators full control over viewpoint transitions can mitigate this issue, it also imposes additional physical and cognitive demands, especially during complex, skill-intensive tasks such as welding [28].

Existing multi-user or multi-view teleoperation systems offer autonomous transitions, but these solutions are not directly applicable to hybrid viewpoint control systems in construction. Rule-based autonomous transitions [17,18,25], which rely on pre-defined rules such as task-related cues, often trigger unintended transitions, resulting in unstable and unreliable user experiences. Learning-based methods [29], such as those using reinforcement learning (RL), also face challenges in defining clear objectives for optimal viewpoint transitions during construction tasks. These challenges have led researchers to explore continuous viewpoint optimization using heuristic rules [16,18–20,23,30] or learning-based methods [14,21,31–33]. However, construction tasks frequently involve unexpected scenarios requiring ad-hoc improvisation [7,34,35], making it nearly impossible to learn viewpoints for every situation. As human-initiated viewpoint transitions are driven by viewing strategies and situational judgements of human experts, we argue that autonomous transitions should imitate these human transition behaviors to ensure reliability and better align with the operator's workflow.

In this work, we propose a viewpoint control mode prediction model that autonomously manages transitions between robot-coupled and decoupled viewpoints during teleoperation in construction tasks. Building on our prior work [25] with rule-based transitions, we extend this by learning from human-initiated transitions collected through a hybrid viewpoint control system with manual switching, enabling the policy model to replicate human-like transition behaviors. We tested our method on simulated welding tasks within a virtual reality (VR) environment, comparing our method to two baselines, namely a standard behavioral cloning (BC) method [36] and a velocity-based transition method [25].

This work expands the research landscape of teleoperation systems for construction tasks by offering the following three contributions. First, we develop an effective learning method based on our transition-guided weighting scheme to exploit sporadic viewpoint transition commands during teleoperation in construction. Second, we systematically evaluate our method against widely used imitation learning baselines and demonstrate its effectiveness through a user study conducted in a real-scale construction site using VR. Third, through a quantitative ablation study, we analyze the effect of individual design choices in our learning method on overall performance.

2. Related work

Our work builds upon previous work in visual perception for teleoperation in construction and viewpoint transition methods for visual perception systems.

2.1. Visual perception for teleoperation in construction

Teleoperation has become integral to a variety of construction applications. It offers adaptability to dynamic environments by combining human decision-making with robotic capabilities [34], while requiring relatively simple technological infrastructure [8]. Most teleoperation systems consist of two major components: actuation and perception [10]. In construction robotics, previous work has focused on improving actuation through enhanced control systems [1] and latency compensation methods [37]. Meanwhile, research in perception has focused on the design of interfaces with visual [38–41], haptic [42,43], and other sensory feedback systems [44].

One of the unique challenges in teleoperation is the dynamic coordination of perception and action [18]. Perception-action coordination depends heavily on the quality of visual feedback provided to the operator, directly influencing their situational awareness—a critical factor for ensuring safety and productivity of remote construction tasks [15,16,19,24,45]. To address this, researchers have explored various viewpoint designs, including robot-coupled, robot-decoupled, and hybrid (dynamic) viewpoints [24].

The robot-coupled viewpoint directly links the operator's perspective to the task robot's movements [24], as shown in Fig. 1A and B. This approach has notably been implemented in orbital camera systems [45,46] and has been shown to outperform standard head-mounted display (HMD) interfaces in perceived usefulness [45]. However, the robot-coupled viewpoint can cause fatigue due to a narrow viewing angle that restricts situational awareness [47]. In our previous work [25], we found that viewpoint control techniques significantly affect performance and user experience during teleoperation in construction. For instance, operators spent 116.7% more time for exploration (excluding lift and welding operations) when using robot-coupled viewpoints instead of decoupled viewpoints.

In contrast, the robot-decoupled viewpoint allows operators to observe the task robot independently from the robot's frame of reference [24], as shown in Fig. 1C and D. This approach, often employing head tracking systems based on vision sensors [48] or HMDs [11], has been adopted in various active vision (AV) systems, including unimanual [49,50] or bimanual teleoperation [13], and mobile robot navigation [51]. However, managing both the camera and the manipulator simultaneously can increase cognitive load, particularly for novice operators.

Hybrid viewpoint aims to balance these trade-offs by allowing operators to switch between coupled and decoupled viewpoints. Such flexibility is particularly beneficial for multi-phase construction tasks. In our previous work [25], we observed that the hybrid viewpoint control technique reduced physical demand and improved welding quality, by enabling operators to mitigate occlusions through viewpoint switching. However, determining the optimal timing for transitions remains a major challenge in hybrid viewpoint control. In our prior study [25], operators often reported frustration with poorly timed robot-initiated transitions and preferred having control over switching. This suggests an important new research area: developing autonomous transition strategies that align with operator intentions while remaining predictable and minimally disruptive to the workflow.

2.2. Viewpoint transition methods for visual perception systems

Remote teleoperation presents challenges in visual perception, including limited field of view (FoV), unfamiliar reference frames, and reduced depth perception [18]. To mitigate these limitations, perception systems offer multiple viewing options, such as multi-view selection with manual or autonomous transitions between viewpoints. In this work, we focus on multi-view systems that allow operators to switch between views, rather than combining multiple perspectives via stitching or merging of multi-user [40,52] or multi-camera [12,53–55] views. However, managing multiple viewpoints can increase cognitive load due to sensory overload [31,56] and misaligned control frames [16], underscoring the necessity of optimal viewpoint selection through transitions.

Previous work has investigated human-initiated transition methods for changing viewpoints or user locations in VR [22,57], selecting optimal cameras for multi-camera teleoperation systems [18,27,28], and switching between robot-coupled and decoupled views [25] or egocentric and exocentric views [58–60]. Human-initiated methods provide operators autonomy for selecting the best preferred viewpoints via buttons [25], robot queries and operator responses [18], and gaze-based controls [28]. However, manual control of camera selection and

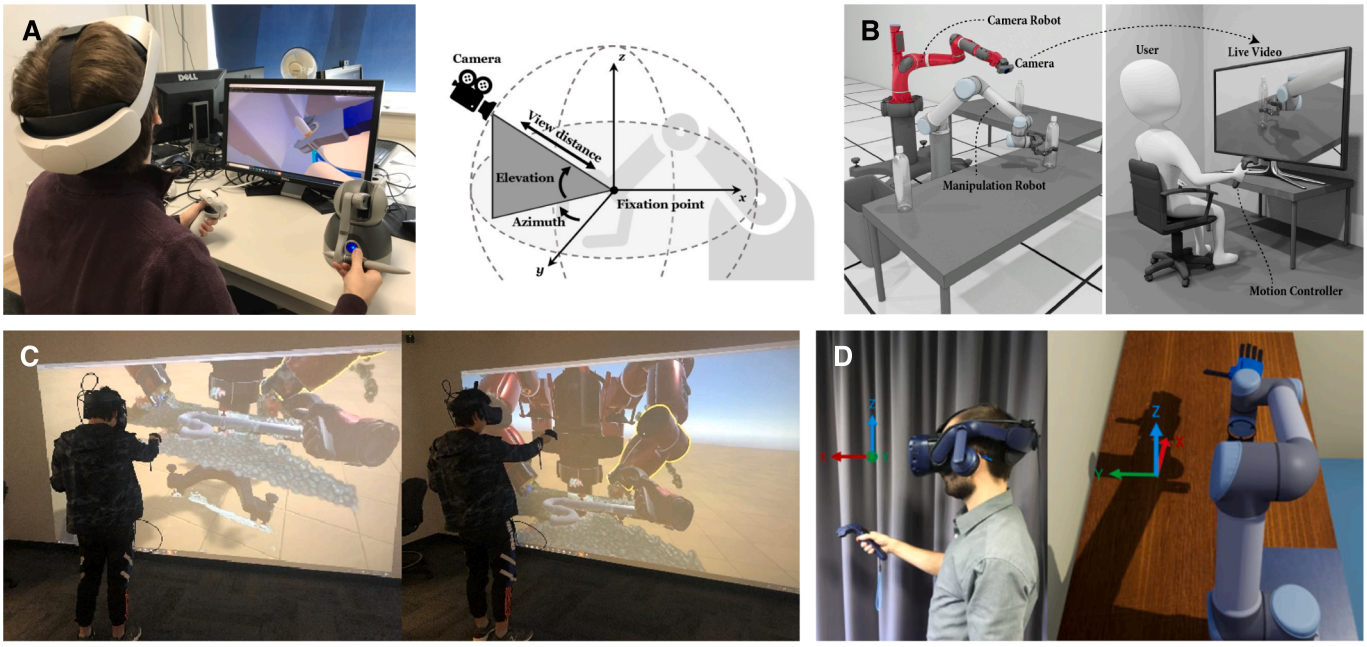


Fig. 1. Examples of viewpoint control designs: (A) Robot-coupled viewpoint with orbital camera [45], (B) robot-coupled viewpoint with autonomous adjustment [19], (C) robot-decoupled viewpoint with HMD head tracking and exocentric view [12], and (D) robot-decoupled viewpoint with HMD head tracking and egocentric view [22].

transition can impose a significant cognitive load, particularly for novice operators [18,61].

To address such challenges of human-initiated transitions, previous work has explored autonomous transitions, using pre-defined rules to avoid obstacles and occlusions [17] and task-specific features, such as end-effector velocity [25] and task error types [18]. However, rule-based transitions often trigger unintended transitions, leading to inconsistent user experiences. Previous work has also investigated learning-based methods, including reinforcement learning (RL)-based policies, to optimize viewpoint transitions by maximizing task-specific rewards such as target tracking performance [29]. Some studies have assessed viewpoint quality by leveraging concepts such as viewpoint entropy [62] and environmental affordance [63,64]. However, defining clear transition objectives or rewards remains difficult, due to the absence of well-established criteria for optimal views or viewpoint control in construction teleoperation tasks. To address these challenges, we propose a viewpoint control mode prediction model that autonomously manages transitions by learning from human-initiated interactions with the viewpoint control system.

Another line of research has proposed the concept of autonomous viewpoint adjustment, in which the viewpoint is continuously optimized rather than selected through discrete transitions. These methods employ heuristics [16,18–20,23,30], domain knowledge [65], RL-based policy [32,33], or learning from human demonstrations [14,21,31] to generate and adjust viewpoints. In particular, recent research has focused on imitation learning-based autonomous policies for dynamic viewpoint control using camera-mounted high-DoF robotic arms, aiming to reduce perceived task load [21] and replicate human-like head movements during remote perception [14]. While effective for simple tasks such as pick-and-place, these methods struggle in complex, skill-intensive tasks where user preferences for viewpoint positioning and timing vary significantly. Construction tasks, in particular, involve unexpected situations that require on-the-fly improvisation [7,34,35], making it nearly impossible to learn viewpoints for all situations. Moreover, inadequate viewpoint control may degrade performance even compared to fully manual robot-decoupled viewpoint control [25].

To this end, instead of a fully autonomous viewpoint adjustment system, we propose a viewpoint control system that autonomously transitions between robot-coupled and decoupled control modes while allowing operators to dynamically adjust viewpoints within each mode. This approach ensures continuous user autonomy over viewpoint, similar to AV systems that enable operators to adjust their viewpoints for improved perspective [9], while providing viewpoint control behavior based on task demands. Overall, we extend prior work by developing a viewpoint control system that integrates both robot-coupled and decoupled viewpoints and provides autonomous transitions through a state transition model learned from human demonstrations.

3. Learning viewpoint control from human-initiated transitions

As described in Section 2.2, instead of developing a fully autonomous viewpoint system that autonomously adjusts camera positions and orientations, we focus on a visual perception system that autonomously transitions between viewpoint control modes. In this way, operators have the autonomy to manually adjust the position and orientation of the viewpoint, while the viewpoint behavior is governed by the current viewpoint control mode, either coupled or decoupled. Accordingly, we define the problem as predicting this binary viewpoint control mode as the policy output, based on current states, such as the proprioceptive states of the robot, taken as the policy input. This section provides an overview of our proposed system, comprising two main components: (1) a teleoperation system with hybrid viewpoint control, in which operators simultaneously teleoperate both a tool-mounted robotic arm and a camera-mounted robotic arm (Section 3.1) and (2) a viewpoint control mode prediction model, which predicts and autonomously switches between viewpoint control modes (Section 3.2).

3.1. Teleoperation system with hybrid viewpoint control

Our teleoperation system is illustrated in Fig. 2. The system consists of a mobile lift with two robotic arms, both controlled by a single operator: a “tool-mounted arm” equipped with an end-of-arm-tool (EOAT) controlled by the operator’s hand movement via a VR controller, and a “camera-mounted arm” equipped with a vision sensor controlled by the operator’s head movement via an HMD. This bimanual teleoperation setup, incorporating additional robotic arms with cameras

Fig. 2. Teleoperation system setup. (A) Mobile lift with two robotic arms, the camera-mounted arm and the tool-mounted arm. (B) The operator teleoperating the system using a VR controller. (C) The operator's HMD view showing the tool-mounted arm and the visual feedback from the camera-mounted arm.

for dynamic visual feedback, builds upon prior telemanipulation systems [18–23], particularly those used in AV systems [9,10,14].

The hybrid viewpoint control allows operators to switch between robot-coupled and decoupled modes. Based on our previous work [25,66], we implement a head motion-based control method for the robot-decoupled viewpoint and a heuristic-based viewpoint optimization method [16–21,23] for the robot-coupled viewpoint. In the robot-coupled viewpoint, the goal pose of the camera-mounted arm is calculated using a weighted-sum nonlinear optimization, which considers three objectives, including avoiding collisions with the tool-mounted robot, positioning the tool-mounted arm's end-effector near the center of the viewpoint, and minimizing the positional difference between the current and desired poses of the camera-mounted robot. Transitions between these two viewpoint control modes are initiated by toggling the grab button on the left-hand VR controller.

As an alternative to manual transitions, we also implement a velocity-based transition method as a baseline, adapted from our prior work [25]. This method infers transition timing based on the end-effector velocity of the tool-mounted arm. Specifically, transitions to the coupled viewpoint control mode are triggered when the end-effector velocity remains below a predefined threshold, indicating fine manipulation (e.g., welding or precise positioning). Conversely, transitions to the decoupled viewpoint control mode occur when the velocity exceeds this threshold, indicating a broader manipulation or exploratory actions. In this study, to minimize abrupt mode-switching, we enforce a duration threshold of 2 s before confirming a viewpoint transition, as in our prior implementation [25]. Similar velocity-based approaches have been commonly employed in previous telemanipulation studies for task phase judgement [67].

For the tool-mounted arm, we employed a direct mapping technique to synchronize the position and orientation of the right-hand VR controller with the tool-mounted arm end effector, enabling direct control over the robot [68]. Additionally, since our system is designed for construction tasks at height, lift control is integrated into the setup. The VR controller's thumbstick is used for horizontal movements (forward/backward and left/right), and pressing the trigger toggles the control mode to enable vertical movement (up/down) of the mobile lift.

3.2. Viewpoint control mode prediction model

Fig. 3 shows the overview of the viewpoint control mode prediction model. In the training phase, we collect a demonstration dataset from multiple human operators during their interactions with the teleoperation system with hybrid viewpoint control. Here, operators transmit transition commands to the system; however, we collect the binary viewpoint control modes directly from the hybrid viewpoint control system as actions, as the goal of our policy is to predict viewpoint control modes. The collected action sequences are then preprocessed using a valid transition filter before being used as a multi-human demonstration dataset for policy training. The purpose of the valid transition filter is to exclude human-initiated transitions deemed invalid, typically caused by noise or user errors. Our empirical

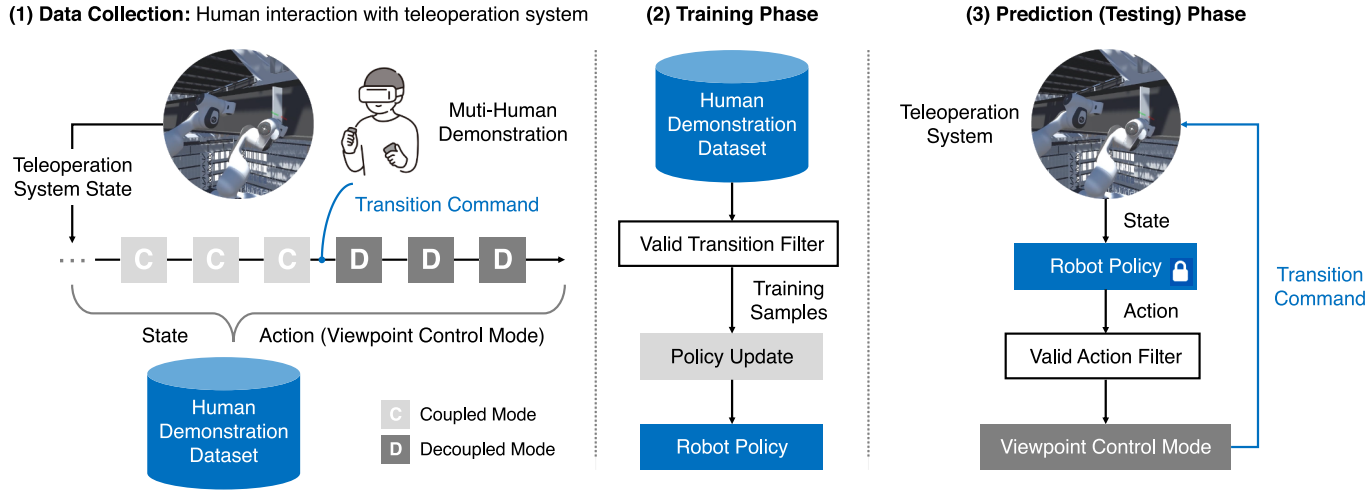


Fig. 3. Overview of the viewpoint control mode prediction model with our learning scheme with human-initiated transitions. The process begins with data collection during human interaction with the teleoperation system, where binary viewpoint control modes are collected. In the training phase, the human demonstration dataset is filtered to remove invalid transitions and used to train a robot policy. In the prediction phase, the trained policy outputs actions, which are post-processed to ensure smooth transitions by replacing short-duration actions with the last valid action.

observations from prior work [66] indicate that human operators revert to the previous viewpoint control mode within 3 s if the newly changed mode is either not preferred or unintentionally triggered. Thus, when a viewpoint control mode fails to persist for a threshold duration of 3 s, the transitions leading to and from this mode change are considered invalid, and the entire sequence of modes within this invalid period is reverted to the last valid mode.

In the prediction (testing) phase, we use the trained robot policy to output binary actions. The action sequences undergo post-processing through valid action filtering to provide smooth transitions for operators. In this work, an action is considered valid if it is sustained for a minimum duration of a threshold time of 1 s, based on our prior studies [66]. Actions that do not satisfy this condition are substituted with the last valid action. Valid actions, binary viewpoint control modes, are then transmitted as final action commands to the robot.

3.2.1. Learning from human-initiated transition commands

We frame our viewpoint state prediction as a sequential decision-making process, formalized using the Markov Decision Process (MDP) framework [69]. An MDP models a stochastic, sequential decision-making process in a fully observable environment [70] as a tuple $\mathcal{M} = < \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma >$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{T}(s'|a,s)$ is the transition model, a probability distribution over the next state s' (i.e., s_{t+1}) given the current state s (i.e., s_t) and action a (i.e., a_t), $\mathcal{R}(s,a)$ is the reward function, and γ is the discount factor, where $\gamma \in [0,1]$.

In the MDP, a deterministic policy π is a mapping from states to actions ($\pi : \mathcal{S} \rightarrow \mathcal{A}$), defining which action to take in a given state [70]. The objective of the MDP is to find a policy π^* which maximizes the expected discounted accumulated reward over an infinite horizon:

$$\pi^* = \underset{\pi}{\text{argmax}} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t)\right]$$

Behavioral cloning (BC) is a commonly used imitation learning method that aims to learn a direct mapping from states to actions, by framing the problem of learning to imitate an expert as a supervised learning problem [71]. BC assumes access to a dataset of expert demonstrations $\mathcal{D} := \left\{\left(s_1^i, a_1^i\right), \left(s_2^i, a_2^i\right), \ldots, \left(s_{T^i}^i, a_{T^i}^i\right)\right\}_{i=1}^{N}$, where T^i is the length of the i-th trajectory and N is the total number of trajectories. Here, a trajectory refers to a sequence of states and actions experienced during a single cycle of interaction between the expert or robot and the environment [70,71]. The goal of BC is to train a policy π parameterized by θ by minimizing the error between the policy's predicted action $\pi_\theta(s)$ and expert action a:

$$\theta^* = \underset{\theta}{\text{argmin}} \mathbb{E}_{(s,a)\sim\mathcal{D}} \|\pi_\theta(s) - a\|_2^2$$

For the system policy, we employ BC-RNN, a state-of-the-art BC algorithm with a recurrent neural network (RNN) backbone [36,72]. Fig. 4 shows the policy architecture of the adopted BC-RNN model. The input to the robot policy consists of a concatenation of several key features: (1) position and orientation of the head-mounted display (HMD), (2) proprioceptive states of the tool-mounted arm, (3) linear velocity of the mobile lift, (4) distances between the tool tip and task targets, and (5) task-specific states. The proprioceptive states of the tool-mounted arm include the position and orientation of the EOAT relative to the robot base, as well as its linear velocity. The task-specific states are represented by four binary indicators: (1) tool trigger state $s^{trigger}$, which indicates whether the tool is activated, (2) lift control state s^{lift}, which indicates whether the mobile lift is in operation, (3) proximity state $s^{proximity}$, which indicates the robot's proximity to the target workspace and (4) contact state $s^{contact}$, which indicates whether the tool is in contact with the workpiece. The long short-term memory (LSTM) network outputs hidden states as well as a value between 0 and 1, which is converted into a binary viewpoint control mode (action) using an action threshold.

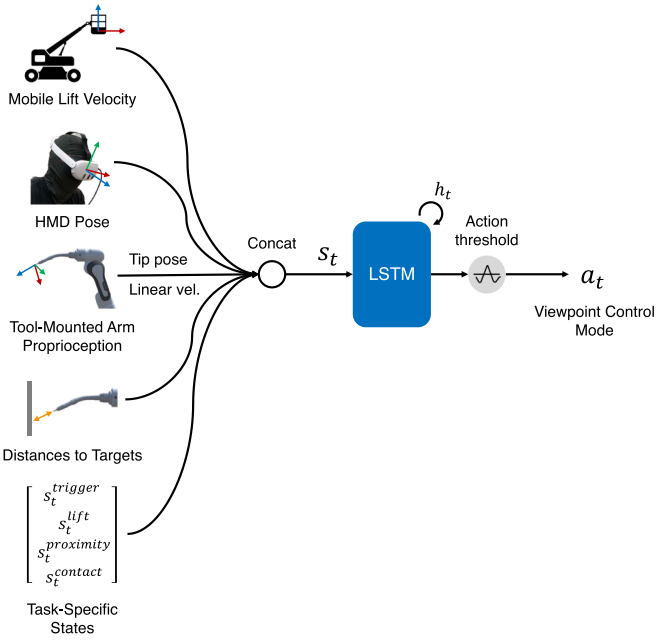


Fig. 4. Policy architecture of the BC-RNN model.

3.2.2. Transition-guided weighting scheme design

In this work, we utilize weighted BC, which aims to improve the reliability of BC by prioritizing high-quality samples for learning through a weight function $w(s,a)$ that depends on the state-action pair [72,73]. In general, the weight function seeks to assign greater weights to desirable state-action pairs and smaller weights to suboptimal behaviors [74]. Consequently, we can modify our objective function of BC as follows:

$$\theta^* = \underset{\theta}{\text{argmin}} \mathbb{E}_{(s,a)\sim\mathcal{D}} \left[w(s,a) \cdot \|\pi_\theta(s) - a\|_2^2\right]$$

Weighted BC has recently become a promising approach for learning policies from multimodal and suboptimal data, particularly in the context of offline RL [72]. Previous work has explored various weighting approaches, such as high-return trajectory weighting [73] and weight functions based on the log ratio of suboptimal and expert behavior densities [75] or exponential advantage [74]. Previous work has also investigated data sampling methods that utilize a previously trained policy on suboptimal demonstrations [76] or apply importance sampling rules to adjust class distributions [72]. In particular, importance sampling has proven to be a simple yet effective approach for human-in-the-loop robot learning methods [72,77], in which weights are designed to leverage critical information around human intervention occurrences.

Building upon the foundation of importance sampling, we propose five weighting schemes for training viewpoint transitions (Fig. 5). Drawing insights from previous studies on intervention-guided weighting [72,77], we hypothesize that timesteps surrounding transitions contain the most valuable information for policy training. The five transition-guided weighting schemes are designed to address challenges in viewpoint control mode prediction, including class imbalance, variability in transition behavior, and most importantly, the sporadic nature of transitions. The details of these schemes, along with the rationale for each, are as follows:

1. Class-balanced sampling: Class imbalance is addressed by undersampling or oversampling the coupled and decoupled modes. The

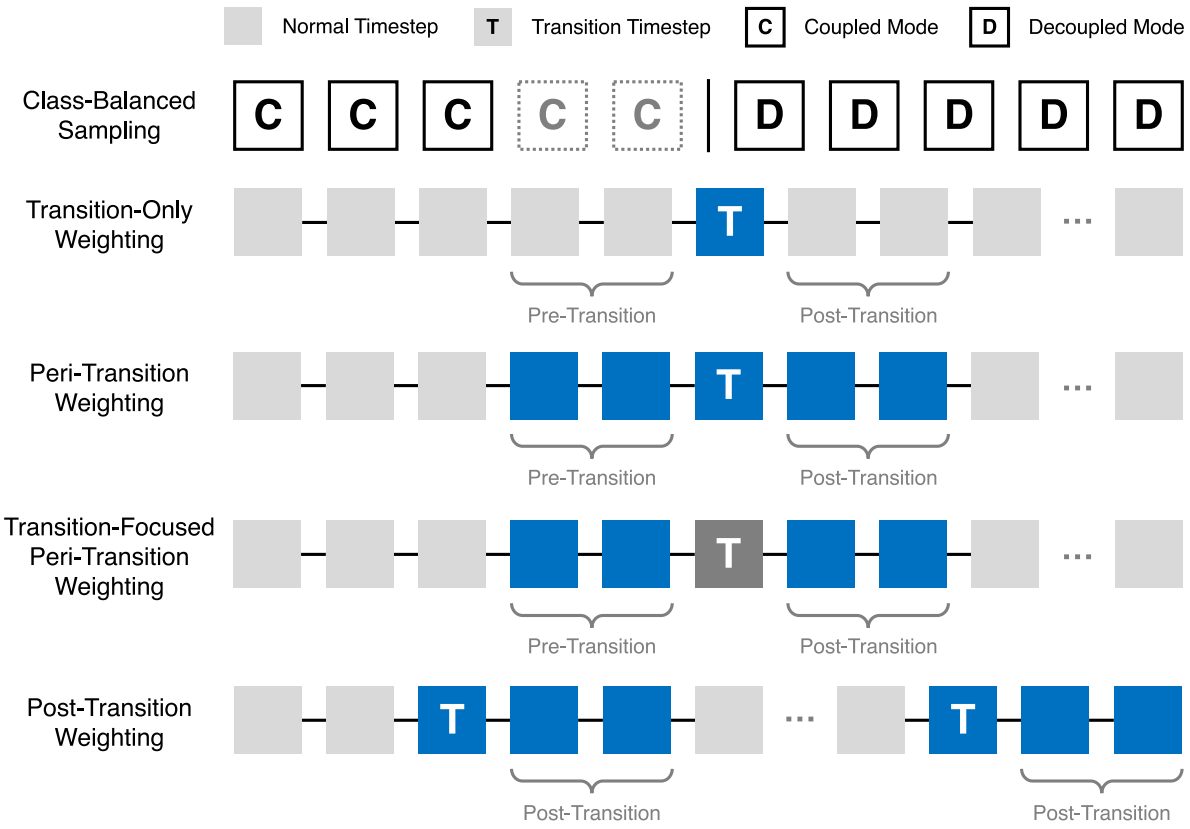


Fig. 5. Weighting schemes for behavioral cloning. Each box represents a single timestep (state-action pair). Grey dotted-line boxes in class-balanced sampling indicate upsampled mode instances used to balance class distributions. Dark grey and blue boxes represent weighted timesteps, with each color corresponding to a distinct uniform weight within the weighting scheme. Light grey boxes represent unweighted timesteps.

aim of policy learning can be viewed as a binary classification task where class balance is crucial.

2. Transition-only weighting: Weights are assigned only to the transition timesteps, while the weights of pre-transition, post-transition, and normal states remain unchanged. As our focus is on predicting the timing of transitions, these transition timesteps contain critical information.
3. Peri-transition weighting: Weights are assigned to the pre-transition, transition, and post-transition timesteps, preserving their relative class distribution, while adjusting the weights of normal timesteps accordingly. Transitions should be considered in the context of pre- and post- transitions, thus should not be oversampled.
4. Transition-focused *peri*-transition weighting: Weights are assigned to the transition timesteps, and the number of pre- and post-transition samples remains the same. The weights of normal timesteps are adjusted accordingly. While pre- and post-transition timesteps provide important contextual information for learning, transition timesteps contain the most critical state data.
5. Post-transition weighting: Weights are assigned to the transition and post-transition timesteps, preserving their relative class distribution, while adjusting the weights of pre-transition and normal timesteps. Since our valid transition filtering method removes invalid transitions, post-transition timesteps contain more important information than pre-transition timesteps, as they may reflect user confidence in the new state.

4. Experimental results and Discussion

In our experiments, we aim to address the following research questions: (1) How effectively does our model predict the viewpoint control mode? (2) How do individual design choices in our learning scheme impact overall performance? This section provides a detailed description of the experimental setup, tasks, and evaluation protocol, followed by experiments and analyses addressing these research questions.

4.1. Experimental setup and tasks

We designed our experimental setup within a VR environment that simulates a virtual industrial hall, reflecting real-world construction settings. The virtual hall included structural elements such as steel columns, beams, roof trusses, and reinforced concrete spot-footings. The experiment utilized a VR system, including a Meta Quest 3 VR headset and Meta Quest Touch Plus controllers. The mobile lift is a four-wheeled lift with a height range of 2.9 to 9.6 m. The two robotic arms are Franka Emika Panda robots, controlled in joint position. We focused on welding as our application scenario, due to its requirement for skill-intensive manipulation and frequent viewpoint adjustments to maintain optimal visibility and task precision [8,25,78–80]. For the experiments, the tool-mounted arm was equipped with a gas metal arc welding (GMAW) tool, commonly referred to as a MIG welding gun, and the camera-mounted arm was equipped with an RGB-D camera to provide real-time visual feedback. In our implementation, the VR controller's orientation and index trigger were aligned with those of a real welding gun, as shown in Fig. 6A. During the welding operation, the operator activated wire feeding by pressing the right-hand VR controller's trigger, mirroring real-world welding gun operation.

We conducted a user study involving 10 participants to collect expert demonstration trajectories. The experimental protocol was approved by the authors' Institutional Review Board (IRB No. 2306/002–002). Participants were tasked with welding 5 stiffeners onto a steel beam, resulting in 10 weld seams per trial at a height of 6.8 m (Fig. 6B). For each weld seam, participants navigate the mobile lift to approach the

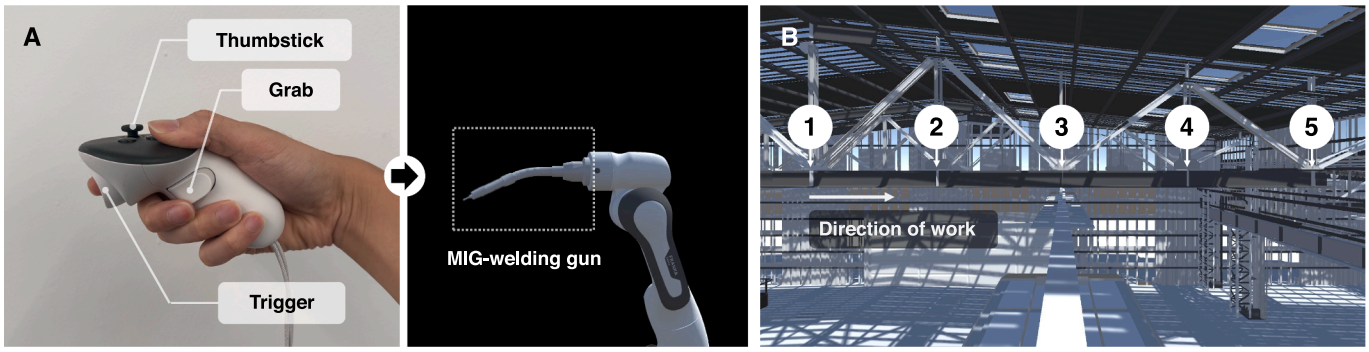


Fig. 6. Experimental setup. (A) Direct mapping of the VR controller and the MIG-welding gun. (B) Locations of the five weld beam stiffeners. Participants welded in sequence from 1 to 5.

weld area, position the tip of the welding gun at the initial point, and perform welding using the vertical up technique. Overall, we collected 10 expert trajectories, with each trajectory representing a single trial by a different participant. The total trajectory length was 32,219, with a total duration of approximately 54 min (10 Hz sampling frequency). Fig. 7 illustrates how human operators utilized the two viewpoint control modes during the welding of a single seam in a sample trial by Participant #2.

4.2. Evaluation protocol

To establish a robust evaluation protocol, we conducted additional experiments with two participants who performed the same tasks using our teleoperation system with manual transitions. The participants were graduate students in civil engineering with prior experience in VR systems. Prior to the experiment, participants watched a brief expert welding video tutorial and completed a guided hands-on VR training session. These participants completed the tasks with an average trial duration of 6.0 min (SD = 0.4) per trajectory. The collected expert trajectories, with a total trajectory length of 7,162, were treated as ground truth for evaluation. These trajectories provide action sequences that serve as reference benchmarks for assessing the model performance.

We evaluate the model performance in two aspects, employing two commonly used metrics in imitation learning [81–83]. First, we evaluate transition timing prediction performance using the F1 score, the harmonic mean of precision (the proportion of true positive transitions out of all predicted positive transitions) and recall (the proportion of true positive transitions out of all actual positive transitions). A transition is considered correct if both the direction and timing are accurate. Here, we set the tolerance for correct transition timing predictions to 3 s, corresponding to the tolerance used in the valid transition filtering. In this work, the transition timing score is selected as the primary evaluation metric, as it aligns with the model's objective to replicate the transition timing of human operators. However, since the transition timing score mainly focuses on the timesteps surrounding transitions, we also evaluate the model's general performance by measuring viewpoint control mode accuracy across the entire sequence. Here, accuracy is defined as the ratio of correct viewpoint control mode predictions to the total number of predictions [71]. In our context, the total number of predictions corresponds to the total number of timesteps.

We compared our method, BC-RNN with *peri*-transition weighting, with a standard BC-RNN (described in Section 3.2) and a velocity-based judgement method from our prior work [25]. Furthermore, we compared two state-of-the-art intervention-guided weighted BC algorithms (also described in Section 3.2), derived from intervention weighted regression (IWR) [77] and Sirius [72]. To implement IWR, we calculated weights for our transition timesteps using the same method applied to human intervention samples in IWR. For Sirius, we adapted its weighting scheme by applying the same target distributions for human interventions, pre-intervention, and demonstration samples to our transition, pre-transition, and normal timesteps, respectively. For all BC-related methods, including standard BC-RNN, IWR, Sirius, and ours, we used the same policy architecture, which includes 2-layer LSTMs, each with 400 hidden dimension and 10 sequence length. The learning hyperparameters are detailed in Table 1.

4.3. Experimental results

Table 2 shows the experimental results from the two participants. The results suggest that our method outperforms the baselines in terms of precision, F1 score for transition timing, and viewpoint control mode accuracy. Specifically, our method significantly outperforms the

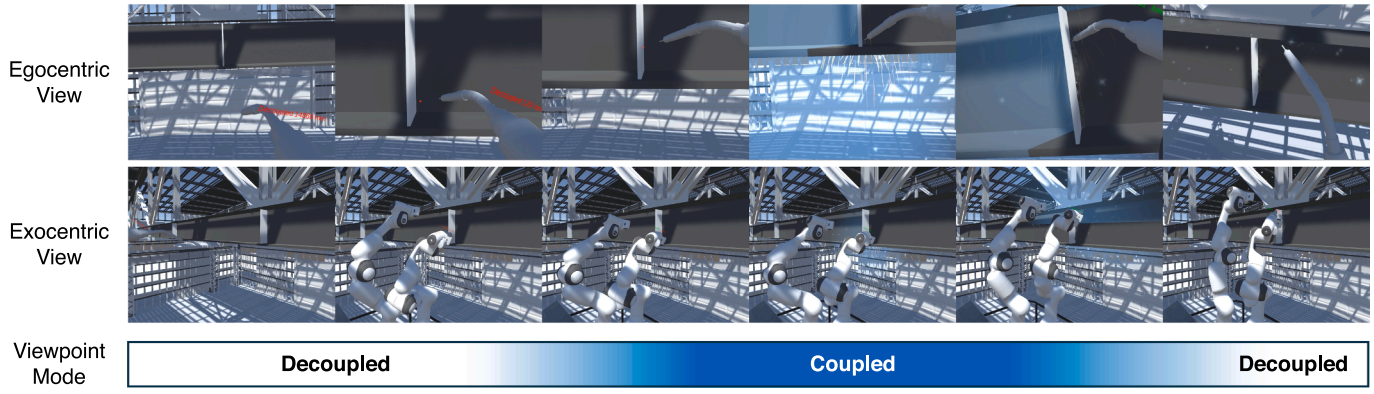


Fig. 7. Egocentric and exocentric views with corresponding viewpoint control modes during a welding trial by Participant #2. The operator navigates the mobile lift to approach the weld area and performs vertical up welding. The robot-coupled viewpoint is used from the moment the welding gun tip is positioned at the starting point until the weld is completed. Afterward, the operator transitions back to the robot-decoupled viewpoint to explore the next target seam.

Table 1
Common hyperparameters.

Hyperparameter	Value
Learning rate	1e-6
Batch size	32
Number of training steps per epoch	500
Number of training epochs	1000
Optimizer	Adam

Table 2
Quantitative results.

Method	Transition timing			Viewpoint control mode
	Precision	Recall	F1 score	Accuracy
Velocity-based transition [25]	0.129	**0.619**	0.174	0.323
BC-RNN [36]	0.571	0.571	0.571	0.882
IWR [77]	0.129	**0.619**	0.204	0.865
Sirius [72]	0.625	0.464	0.531	0.875
Ours	**0.775**	0.536	**0.633**	**0.921**

velocity-based transition baseline, highlighting the effectiveness of using imitation learning to replicate human transition behaviors in viewpoint control. However, the velocity-based transition method yields higher recall due to its sensitivity in detecting transitions (see Fig. 8). Importantly, our method outperforms the standard BC-RNN, IWR, and Sirius across nearly all metrics, accurately predicting most transitions except for a few transition commands of relatively short duration. We attribute this improvement in transition timing prediction to our *peri*-transition weighting scheme, which enables the model to learn from the full temporal context, including pre-transition, transition, and post-transition timesteps. Interestingly, the results also show that our weighting scheme enhances viewpoint control mode accuracy, suggesting that transition-related samples are particularly beneficial for policy training, not only for transition timing prediction but also for general viewpoint control mode prediction, compared to non-related samples.

Furthermore, we conducted principal component analysis (PCA) on the hidden states of the LSTMs for our model and two weighted BC baselines (IWR and Sirius), and visualized the results in 2D, as shown in Fig. 9. We observed a U-shaped trajectory across all methods, with each numbered vertex corresponding to a specific task phase. In particular, the *peri*-transition clusters highlighted in light blue circles appear more concentrated in our model compared to the baselines, indicating that our model more consistently predicts transitions around specific task phases. While some predicted coupled states appear near decoupled states in the PCA space, this can be attributed to dimensionality reduction and prediction noise near the action threshold.

4.4. *Effect of weighting scheme*

We compare the effectiveness of the transition-guided weighting schemes described in Section 3.2. For each weighting scheme, we conducted an exhaustive search to select the optimal weights for pre-transition, transition, post-transition, and normal timesteps that yield the best transition timing score. As shown in Fig. 10 (Left), all of our transition-guided weighting schemes achieve better performance compared to the standard BC-RNN (Base). Among the weighting schemes, the *peri*-transition scheme, which assigns weights to the pre-transition, transition, and post-transition timesteps, achieves the highest performance in transition timing prediction, improving the F1 score of the standard BC-RNN by 11%. The *peri*-transition weighting also

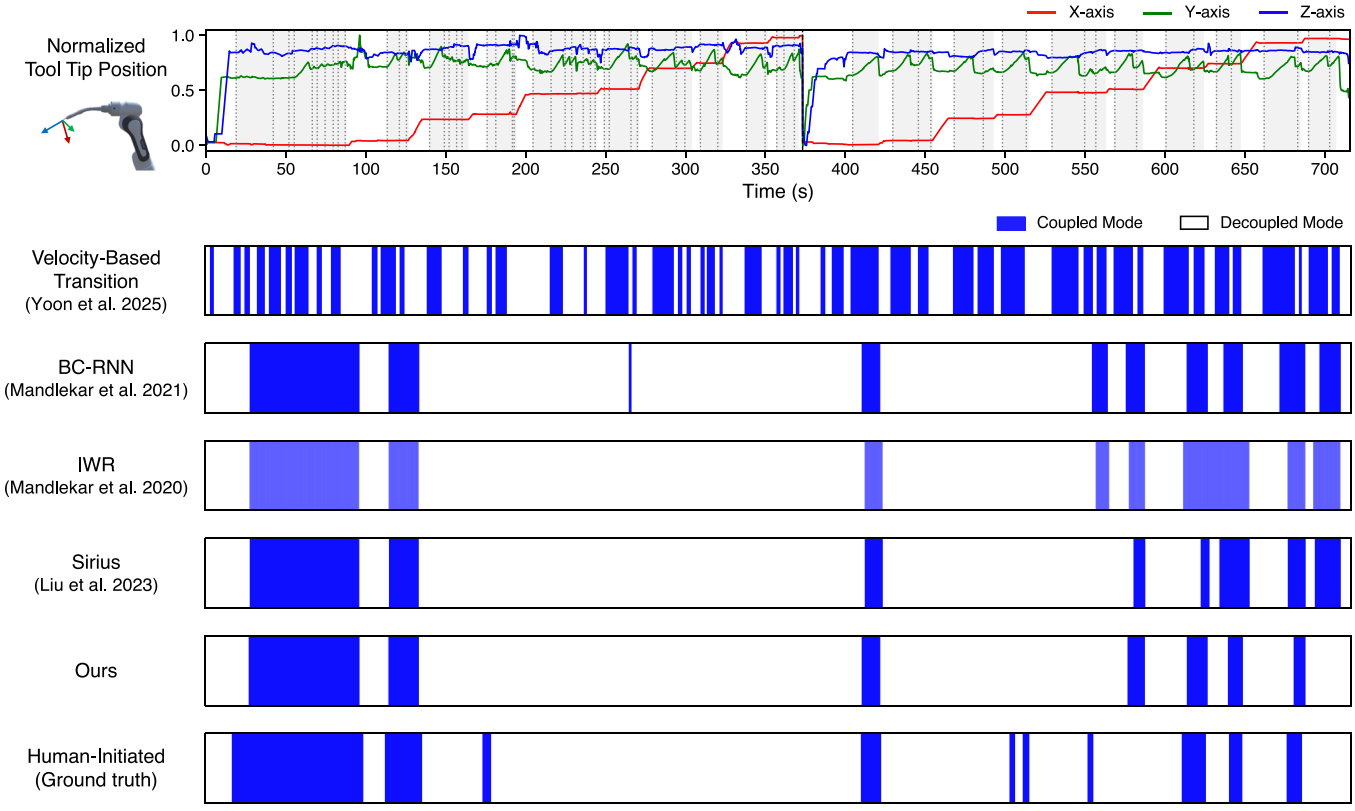


Fig. 8. (Top) Normalized tool tip position of the welding robot over the task horizon. Grey dotted lines mark the timesteps corresponding to the activation of the welding gun trigger. Lightgrey areas represent the duration for each weld seam. (Bottom) Comparison of viewpoint control mode predictions across methods: velocity-based transition, policy rollout of a standard BC-RNN, weighted BC baselines (IWR and Sirius), our method, and ground truth human-initiated transitions.

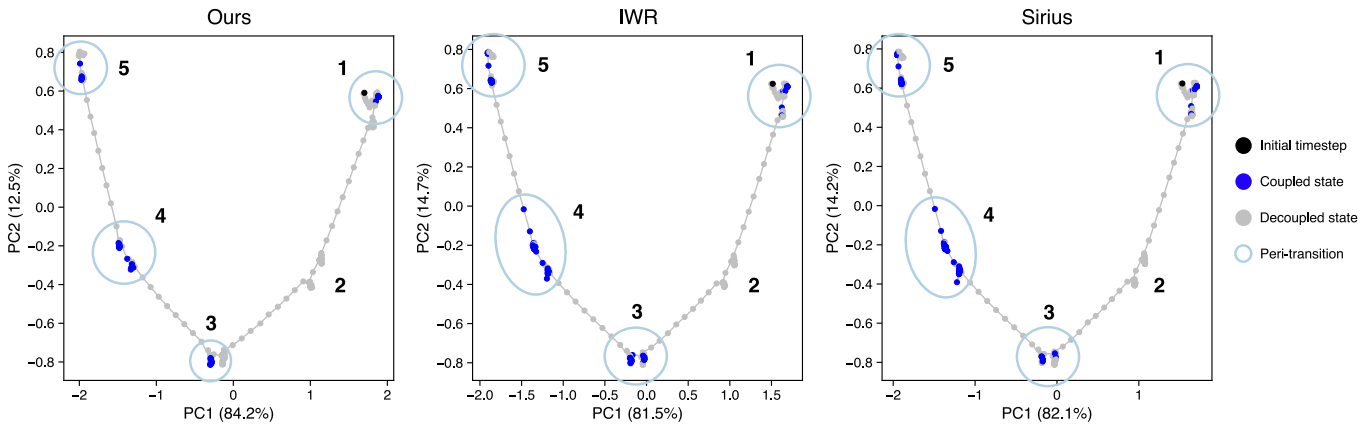


Fig. 9. PCA visualization of the LSTM hidden states for our model and two weighted BC baselines. The x- and y-axes indicate the first two principal components, PC1 and PC2 respectively.

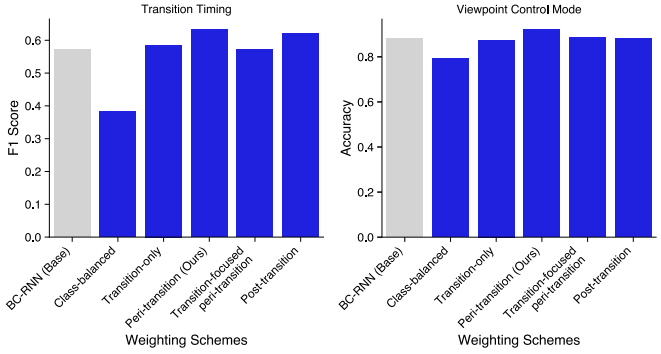


Fig. 10. Ablation on weighting scheme design. (Left) Among all weighting schemes, *peri*-transition weighting achieves the highest transition timing F1 score. (Right) Peri-transition weighting also demonstrates the highest viewpoint control mode accuracy compared to the other schemes.

demonstrates the highest performance in viewpoint control mode accuracy, as shown in Fig. 10 (Right).

When comparing *peri*-transition weighting with other schemes, we identified several differences that may explain its relative effectiveness. First, class-balanced sampling was ineffective, as it does not account for the dynamic nature of viewpoint transitions. Interestingly, contrary to our intuition, transition-based weighting schemes (i.e., transition-only and transition-focused *peri*-transition weighting), which assign higher weights to transition timesteps compared to others, are less effective in predicting transition timing. This suggests that although transition timesteps may represent distinctive states in hybrid viewpoint control, they are inherently unstable and contain noise, making it difficult for the model to generalize and potentially leading to less robust policies for managing autonomous transitions. Rather than oversampling transition timesteps, we conclude that treating pre- and post-transition timesteps with equal importance is more effective in terms of predicting both transition timing and viewpoint mode. While post-transition weighting demonstrates competitive performance in predicting transition timing, it struggles to generalize across the entire sequence, as pre-transition timesteps contain equally valuable contextual information as post-transition timesteps.

We also investigate the effects of the target distribution for *peri*-transition timesteps. Following the rule of importance sampling, we set the weight value for *peri*-transition by dividing the target distribution by the original distribution of training samples [72]. As shown in Fig. 11, both metrics reach their highest values at a target distribution of 0.2, in which weights are set to 1.2 and 0.6 for *peri*-transition and normal states, respectively.

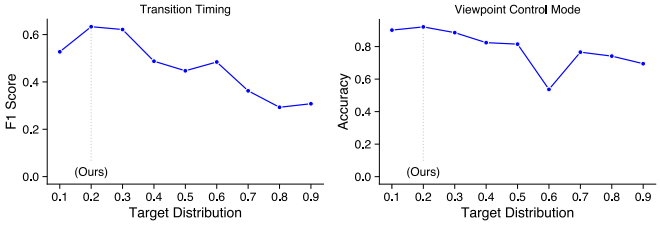


Fig. 11. Ablation on target distribution for *peri*-transition weighting. The highest transition timing F1 score and viewpoint control mode accuracy are observed at a target distribution of 0.2.

4.5. Ablation study

We conducted an ablation study to examine the impact of different observation space configurations on the performance of our policy model. Specifically, we analyzed the effects of removing the HMD pose, tool tip pose, tool tip linear velocity, task states, distances to task targets, or linear velocity of the mobile lift. As shown in Fig. 12, manipulating the observation space led to degradation in policy performance, suggesting that all examined observation spaces are critical for learning the transition task. In particular, removing the HMD pose results in a relative performance drop of 92% and 18% in transition timing score and viewpoint control mode accuracy, respectively.

We also performed an ablation study to examine the effects of (1) filtering strategies, (2) our learning algorithm, and (3) the threshold for binary action output in our policy. First, we evaluated the individual effects of the valid transition filter and valid action filter by removing

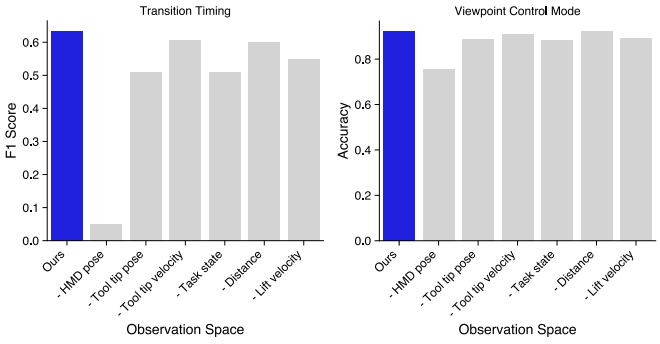


Fig. 12. Ablation on observation space. Removing any features from the observation space results in clear drops in transition timing and viewpoint control performance.

each filtering strategy from our learning scheme. As shown in Fig. 13, our method with both filters outperforms methods without either filter on both metrics. This demonstrates that filtering out invalid transitions is critical for training a consistent and robust policy, while filtering binary action outputs ensures smooth viewpoint control without noisy or inappropriate transitions that may potentially disrupt teleoperation tasks. Next, we evaluated the impact of our learning algorithm by (1) removing the RNN (LSTM) network component, (2) replacing the 2-layer LSTMs with 2-layer gated recurrent units (GRUs), (3) substituting the deterministic policy with a stochastic policy, specifically a Gaussian mixture model (GMM), or (4) replacing the 2-layer LSTMs with 6-layer transformers. As shown in Fig. 13, the BC-RNN used in our policy outperforms all alternatives. In particular, our BC-RNN outperforms the transformer-based BC [84], a current state-of-the-art BC, with a 63% higher transition timing F1 score and a 10% higher viewpoint mode accuracy. Last, we examined the effect of varying the action threshold and observed that a threshold value of 0.5 yields the highest performance for both metrics, as shown in Fig. 14.

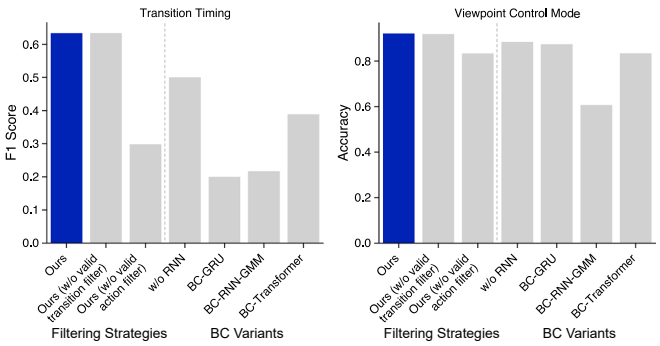


Fig. 13. Ablation on filtering strategies and BC variant learning algorithms. Our method, incorporating both valid transition and action filters, improves transition timing prediction, outperforming methods without the filters. Among the BC variants, the BC-RNN model achieves the highest transition timing score and viewpoint control mode accuracy.

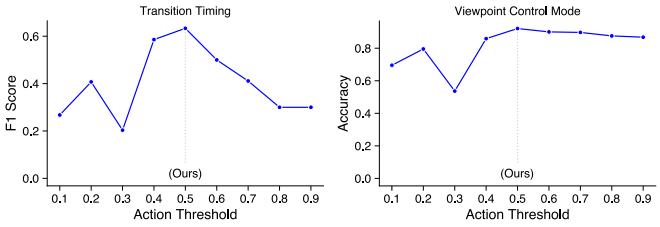


Fig. 14. Ablation on action threshold. Transition timing and viewpoint control performance are optimized with a threshold value of 0.5.

4.6. Limitations and future work

Finally, we discuss the limitations and future directions of this work. First, we evaluated the performance of our policy on expert trajectories with human-initiated transitions, rather than rollouts from our policy. Although this evaluation protocol was designed to establish a reliable ground truth—given that the primary goal of our model is to closely replicate human transition commands—we believe that future work would benefit from qualitative evaluations in real-world scenarios in which our autonomous state transition policy is deployed in a viewpoint control system. Second, participants involved in our user study were novices without prior experience in welding, and thus our findings may not reflect expert behavior during welding tasks. However, this participant group enabled us to minimize inconsistent or biased task execution, which may arise in a user study involving expert welders with varying experience levels. Future work should investigate how different levels of welding expertise impact the generalizability of our policy. Third, for practical implementation, the use of a valid action filter may introduce time delays during teleoperation. While we believe that this delay can be mitigated by reducing the threshold duration to less than 1 s, we suggest that future work explore more rigorous and effective action filtering methods, such as learning-based binary action classification, to address this issue. Finally, our evaluation included only two expert trajectories (total duration of 11.9 min). However, we argue that our telerobotic welding tasks involve extensive activities, including lift operation, exploration, and welding manipulation, resulting in evaluation trajectories that are even 7+ times longer than those in previous long-horizon tasks [85]. A large-scale user study involving diverse skill-intensive tasks and diverse scenarios, such as varying heights and initial robot configurations, would further enhance the model's robustness and generalizability by expanding the training dataset and evaluating its adaptability to broader task conditions and user variability.

5. Conclusion

Hybrid viewpoint control systems for dynamic visual perception provide significant advantages by enabling transitions between robot-coupled and decoupled viewpoints during teleoperated construction tasks. However, determining the optimal timing for these transitions remains a challenge, as existing autonomous transition approaches are not directly applicable to hybrid viewpoint control in construction. In this study, we propose a viewpoint control mode prediction model that autonomously manages viewpoint transitions by learning from human interactions with our teleoperation system. Our user evaluation demonstrates the effectiveness of our learning-based approach in replicating human-initiated viewpoint transitions, including both transition timing and viewpoint control behaviors. Moreover, our results show that assigning weights to *peri*-transition timesteps leads to 11% and 19% improvements in transition timing F1 score over the standard BC and weighted BC algorithms, respectively. Our work offers valuable insights into visual perception systems for teleoperation in construction, particularly in supporting autonomous management of human-like viewpoint transitions.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT in order to improve language and readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

CRediT authorship contribution statement

Sungboo Yoon: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Moonseo Park:** Supervision, Funding acquisition. **Changbum R. Ahn:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the Korea Agency for Infrasdtructure Technology Advancement (KAIA) grant funded by the Ministry of Land,

Infrastructure and Transport (Grant RS-2025-11802969) and the Institute of Construction and Environmental Engineering at Seoul National University.

Data availability

Data will be made available on request.

References

[1] H.J. Lee, S. Brell-Cokcan, Cartesian coordinate control for teleoperated construction machines, Construction Robotics 5 (2021) 1–11, https://doi.org/10.1007/s41693-021-00055-y.

[2] K.H. Koh, M. Farhan, K.P.C. Yeung, D.C.H. Tang, M.P.Y. Lau, P.K. Cheung, K.W. C. Lai, Teleoperated service robotic system for on-site surface rust removal and protection of high-rise exterior gas pipes, Autom. Constr. 125 (2021) 103609, https://doi.org/10.1016/j.autcon.2021.103609.

[3] Y.-P. Su, X.-Q. Chen, C. Zhou, L.H. Pearson, C.G. Pretty, J.G. Chase, Integrating Virtual, mixed, and Augmented reality into Remote Robotic applications: a Brief Review of Extended Reality-Enhanced Robotic Systems for Intuitive Telemanipulation and Telemanufacturing Tasks in Hazardous Conditions, NATO Adv, Sci. Inst. Ser. E Appl. Sci. 13 (2023) 12129, https://doi.org/10.3390/app132212129.

[4] K. Duan, Z. Zou, T.Y. Yang, Training of construction robots using imitation learning and environmental rewards, Comput. Aided Civ. Inf. Eng. (2024), https://doi.org/10.1111/mice.13394.

[5] R. Li, Z. Zou, Enhancing construction robot learning for collaborative and long-horizon tasks using generative adversarial imitation learning, Adv. Eng. Inf. 58 (2023) 102140, https://doi.org/10.1016/j.aei.2023.102140.

[6] Y. Li, S. Liu, M. Wang, S. Li, J. Tan, Teleoperation-driven and keyframe-based generalizable imitation learning for construction robots, J. Comput. Civ. Eng. 38 (2024), https://doi.org/10.1061/jccee5.cpeng-5884.

[7] S. Yoon, M. Park, C.R. Ahn, LaserDex: Improvising Spatial Tasks using Deictic Gestures and Laser Pointing for Human–Robot Collaboration in Construction, J. Comput. Civ. Eng. 38 (2024) 04024012, https://doi.org/10.1061/JCCEE5.CPENG-5715.

[8] H. Hu, A. Song, L. Wei, J. Mao, Development of a virtual reality-based teleoperated welding robot system for enhanced safety and efficiency, in: Proceedings of the 2024 4th International Conference on Robotics and Control Engineering, ACM, New York, NY, USA, 2024: pp. 77–82. DOI: 10.1145/3674746.3674758.

[9] I. Chuang, A. Lee, D. Gao, I. Soltani, Active vision might be all you need: Exploring active vision in bimanual robotic manipulation, ArXiv [Cs.RO] (2024). http://arxiv.org/abs/2409.17435.

[10] X. Cheng, J. Li, S. Yang, G. Yang, X. Wang, Open-TeleVision: Teleoperation with immersive active visual feedback, ArXiv [Cs.RO] (2024). http://arxiv.org/abs/2407.01512 (accessed July 29, 2024).

[11] Y. Chen, L. Sun, M. Benallegue, R. Cisneros, R.P. Singh, K. Kaneko, A. Tanguy, G. Caron, K. Suzuki, A. Kheddar, F. Kanehiro, Enhanced Visual Feedback with Decoupled viewpoint Control in Immersive Humanoid Robot Teleoperation using SLAM, ArXiv [Cs.RO] (2022). http://arxiv.org/abs/2211.01749.

[12] T. Zhou, Q. Zhu, J. Du, Intuitive robot teleoperation for civil engineering operations with virtual reality and deep learning scene reconstruction, Adv. Eng. Inf. 46 (2020) 101170, https://doi.org/10.1016/j.aei.2020.101170.

[13] C. Loconsole, D. Leonardis, A. Frisoli, An augmented-reality-assisted immersive system for robotic arms teleoperation, in: Proceedings of the 6th International Conference on Intelligent Human Systems Integration (IHSI 2023) Integrating People and Intelligent Systems, February 22–24, 2023, Venice, Italy, AHFE International, 2023. DOI: 10.54941/ahfe1002831.

[14] B. Sen, M. Wang, N. Thakur, A. Agarwal, P. Agrawal, Learning to look around: Enhancing teleoperation and learning with a human-like actuated neck, ArXiv [Cs.RO] (2024). http://arxiv.org/abs/2411.00704.

[15] J.S. Lee, Y. Ham, H. Park, J. Kim, Challenges, tasks, and opportunities in teleoperation of excavator toward human-in-the-loop construction automation, Autom. Constr. 135 (2022) 104119, https://doi.org/10.1016/j.autcon.2021.104119.

[16] E. Senft, M. Hagenow, P. Praveena, R. Radwin, M. Zinn, M. Gleicher, B. Mutlu, A Method for Automated Drone Viewpoints to support Remote Robot Manipulation, in: In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2022, pp. 7704–7711, https://doi.org/10.1109/IROS47612.2022.9982063.

[17] M. Kamezaki, M. Miyata, S. Sugano, Video presentation based on multiple-flying camera to provide continuous and complementary images for teleoperation, Autom. Constr. 159 (2024) 105285, https://doi.org/10.1016/j.autcon.2024.105285.

[18] A. Valiton, H. Baez, N. Harrison, J. Roy, Z. Li, Active Telepresence Assistance for Supervisory Control: a User Study with a Multi-Camera Tele-Nursing Robot, in: In: 2021 IEEE International Conference on Robotics and Automation (ICRA), 2021, pp. 3722–3727, https://doi.org/10.1109/ICRA48506.2021.9561361.

[19] D. Rakita, B. Mutlu, M. Gleicher, An Autonomous Dynamic Camera Method for Effective Remote Teleoperation, in: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, Association for Computing Machinery, New York, NY, USA, 2018: pp. 325–333. DOI: 10.1145/3171221.3171279.

[20] D. Rakita, B. Mutlu, M. Gleicher, Remote telemanipulation with adapting viewpoints in visually complex environments, Science and Systems XV, Robotics, 2019 https://par.nsf.gov/biblio/10104548.

[21] R. Jia, L. Yang, Y. Cao, C.K. Or, W. Wang, J. Pan, Learning autonomous viewpoint adjustment from human demonstrations for telemanipulation, ACM Trans. Hum. Robot Interact. (2024), https://doi.org/10.1145/3660348.

[22] A. Naceri, D. Mazzanti, J. Bimbo, Y.T. Tefera, D. Prattichizzo, D.G. Caldwell, L. S. Mattos, N. Deshpande, The Vicarios Virtual reality interface for remote robotic teleoperation, J. Intell. Rob. Syst. 101 (2021), https://doi.org/10.1007/s10846-021-01311-7.

[23] L. Chen, A. Naceri, A. Swikir, S. Hirche, S. Haddadin, Autonomous and teleoperation control of a drawing robot avatar, ArXiv [Cs.RO] (2024). http://arxiv.org/abs/2407.20156.

[24] X. Wang, L. Shen, L.-H. Lee, A systematic review of XR-based remote human-robot interaction systems, ArXiv [Cs.HC] (2024). http://arxiv.org/abs/2403.11384.

[25] S. Yoon, M. Park, C.R. Ahn, Comparing dynamic viewpoint control techniques for teleoperated robotic welding in construction, Autom. Constr. 172 (2025) 106053, https://doi.org/10.1016/j.autcon.2025.106053.

[26] A. Moniruzzaman, D. Rassau, S.M.S. Chai, Islam, Teleoperation methods and enhancement techniques for mobile robots: a comprehensive survey, Rob. Auton. Syst. 150 (2022) 103973, https://doi.org/10.1016/j.robot.2021.103973.

[27] T.-C. Lin, A. Unni Krishnan, Z. Li, Perception-motion coupling in active telepresence: Human behavior and teleoperation interface design, ACM Trans. Hum. Robot Interact. 12 (2023) 1–24, https://doi.org/10.1145/3571599.

[28] P. Praveena, L. Molina, Y. Wang, E. Senft, B. Mutlu, M. Gleicher, Understanding control frames in multi-camera robot telemanipulation, in: In: 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2022, pp. 432–440, https://doi.org/10.1109/hri53351.2022.9889543.

[29] A. Sharma, S. Anand, S.K. Kaul, Intelligent camera selection decisions for target tracking in a camera network, in: In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, https://doi.org/10.1109/wacv51458.2022.00314.

[30] Y.-H. Su, K. Huang, B. Hannaford, Multicamera 3D reconstruction of dynamic surgical cavities: Autonomous optimal camera viewpoint adjustment, In: International Symposium on Medical Robotics (ISMR) IEEE 2020 (2020), https://doi.org/10.1109/ismr48331.2020.9312951.

[31] H. Liu, R. Komatsu, S. Nakashima, H. Hamada, N. Matsuhira, H. Asama, A. Yamashita, Viewpoint selection for the efficient teleoperation of a robot arm using reinforcement learning, IEEE Access 11 (2023) 119647–119658, https://doi.org/10.1109/access.2023.3327826.

[32] M. Lodel, B. Brito, Á. Serra-Gómez, L. Ferranti, R. Babuška, J. Alonso-Mora, Where to look next: Learning viewpoint recommendations for informative trajectory planning, ArXiv [Cs.RO] (2022). http://arxiv.org/abs/2203.02381.

[33] Y.-H. Su, K. Huang, B. Hannaford, Multicamera 3D viewpoint adjustment for robotic surgery via deep reinforcement learning, J. Med. Robot. Res. 06 (2021) 2140003, https://doi.org/10.1142/s2424905x21400031.

[34] S. Yoon, Y. Kim, M. Park, C.R. Ahn, Effects of Spatial Characteristics on the Human–Robot Communication using Deictic Gesture in Construction, J. Constr. Eng. Manag. 149 (2023) 04023049, https://doi.org/10.1061/JCEMD4.COENG-12997.

[35] S. Yoon, J. Park, M. Park, C.R. Ahn, A Deictic Gesture-Based Human-Robot Interface for In Situ Task Specification in Construction, in: Computing in Civil Engineering 2023, 2024: pp. 445–452. DOI: 10.1061/9780784485224.054.

[36] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, R. Martín-Martín, What matters in learning from offline human demonstrations for robot manipulation, ArXiv [Cs.RO] (2021). http://arxiv.org/abs/2108.03298.

[37] J. Du, W. Vann, T. Zhou, Y. Ye, Q. Zhu, Sensory manipulation as a countermeasure to robot teleoperation delays: system and evidence, Sci. Rep. 14 (2024) 4333, https://doi.org/10.1038/s41598-024-54734-1.

[38] K. Duan, Z. Zou, Morphology agnostic gesture mapping for intuitive teleoperation of construction robots, Adv. Eng. Inf. 62 (2024) 102600, https://doi.org/10.1016/j.aei.2024.102600.

[39] R. Ding, M. Cheng, Z. Han, F. Wang, B. Xu, Human-machine interface for a master-slave hydraulic manipulator with vision enhancement and auditory feedback, Autom. Constr. 136 (2022) 104145, https://doi.org/10.1016/j.autcon.2022.104145.

[40] D. Liu, J. Kim, Y. Ham, Multi-user immersive environment for excavator teleoperation in construction, Autom. Constr. 156 (2023) 105143, https://doi.org/10.1016/j.autcon.2023.105143.

[41] P. Xia, H. You, J. Du, Visual-haptic feedback for ROV subsea navigation control, Autom. Constr. 154 (2023) 104987, https://doi.org/10.1016/j.autcon.2023.104987.

[42] T. Zhou, P. Xia, Y. Ye, J. Du, Embodied Robot Teleoperation based on High-Fidelity Visual-Haptic Simulator: Pipe-Fitting example, J. Constr. Eng. Manag. 149 (2023) 04023129, https://doi.org/10.1061/JCEMD4.COENG-13916.

[43] C. Brosque, E.G. Herrero, Y. Chen, M.A. Fischer, Collaborative Welding and Joint Sealing Robots with Haptic Feedback, in: In: 2021 Proceedings of the 38th ISARC, 2021, https://doi.org/10.22260/ISARC2021/0003.

[44] H. Nagano, H. Takenouchi, N. Cao, M. Konyo, S. Tadokoro, Tactile feedback system of high-frequency vibration signals for supporting delicate teleoperation of construction robots, Adv. Rob. 34 (2020) 730–743, https://doi.org/10.1080/01691864.2020.1769725.

[45] S. Kuitert, J. Hofland, C.J.M. Heemskerk, D.A. Abbink, L. Peternel, Orbital Head-Mounted Display: a Novel Interface for viewpoint Control during Robot Teleoperation in Cluttered Environments, in: In: 2023 IEEE/RSJ International

Conference on Intelligent Robots and Systems (IROS), 2023, pp. 1–7, https://doi.org/10.1109/IROS55552.2023.10341733.
[46] M. Talha, R. Stolkin, Preliminary Evaluation of an Orbital Camera for Teleoperation of Remote Manipulators, in: In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019, pp. 2754–2761, https://doi.org/10.1109/IROS40897.2019.8968218.
[47] S.N. Young, R.J. Lanciloti, J.M. Peschel, The Effects of Interface views on performing Aerial Telemanipulation Tasks using Small UAVs, Int. J. Soc. Robot. 14 (2022) 213–228, https://doi.org/10.1007/s12369-021-00783-9.
[48] S. Rahnamaei, S. Sirouspour, Automatic viewpoint planning in teleoperation of a mobile robot, J. Intell. Rob. Syst. 76 (2014) 443–460, https://doi.org/10.1007/s10846-014-0028-7.
[49] L.S. Yim, Q.T.N. Vo, C.-I. Huang, C.-R. Wang, W. McQueary, H.-C. Wang, H. Huang, L.-F. Yu, WFH-VR: Teleoperating a robot arm to set a dining table across the globe via virtual reality, in: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2022. DOI: 10.1109/iros47612.2022.9981729.
[50] Y. Zhu, K. Fusano, T. Aoyama, Y. Hasegawa, Intention-reflected predictive display for operability improvement of time-delayed teleoperation system, ROBOMECH J. 10 (2023) 1–11, https://doi.org/10.1186/s40648-023-00258-8.
[51] P. Stotko, S. Krumpen, M. Schwarz, C. Lenz, S. Behnke, R. Klein, M. Weinmann, A VR System for Immersive Teleoperation and Live Exploration with a Mobile Robot, in: In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019, https://doi.org/10.1109/iros40897.2019.8968598.
[52] K.A. Szczurek, R.M. Prades, E. Matheson, J. Rodriguez-Nogueira, M. Di Castro, Multimodal Multi-User mixed reality Human–Robot Interface for Remote Operations in Hazardous Environments, IEEE Access 11 (2023) 17305–17333, https://doi.org/10.1109/ACCESS.2023.3245833.
[53] D. Wei, B. Huang, Q. Li, Multi-view merging for robot teleoperation with virtual reality, IEEE Robot. Autom. Lett. 6 (2021) 8537–8544, https://doi.org/10.1109/lra.2021.3109348.
[54] S. Xu, S. Moore, A. Cosgun, Shared-control robotic manipulation in virtual reality, In: International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) IEEE 2022 (2022) 1–6, https://doi.org/10.1109/hora55278.2022.9800046.
[55] J. Betancourt, B. Wojtkowski, P. Castillo, I. Thouvenin, Exocentric control scheme for robot applications: an immersive virtual reality approach, IEEE Trans. Vis. Comput. Graph. 29 (2023) 3392–3404, https://doi.org/10.1109/TVCG.2022.3160389.
[56] H. Yanco, J. Drury, J. Scholtz, Beyond usability evaluation: Analysis of human-robot interaction at a major robotics competition, Hum.-Comput. Interact. 19 (2004) 117–149, https://doi.org/10.1207/s15327051hci1901&2_6.
[57] J.-I. Lee, P. Asente, W. Stuerzlinger, Designing viewpoint transition techniques in multiscale virtual environments, In: IEEE Conference Virtual Reality and 3D User Interfaces (VR) IEEE 2023 (2023) 680–690, https://doi.org/10.1109/vr55154.2023.00083.
[58] X. Wang, J. Zhu, A mixed reality based teleoperation interface for mobile robot, in: Mixed Reality and Human-Robot Interaction, Springer, Netherlands, Dordrecht, 2011, pp. 77–93, https://doi.org/10.1007/978-94-007-0582-1_5.
[59] F. Ferland, F. Pomerleau, C.T. Le Dinh, F. Michaud, Egocentric and exocentric teleoperation interface using real-time, 3D video projection, in: Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction, ACM, New York, NY, USA, 2009. DOI: 10.1145/1514095.1514105.
[60] A. Abdullah, R. Chen, I. Rekleitis, M.J. Islam, Ego-to-Exo: Interfacing third person visuals from egocentric views in real-time for improved ROV teleoperation, ArXiv [Cs.RO] (2024). http://arxiv.org/abs/2407.00848.
[61] Z. Li, P. Moran, Q. Dong, R.J. Shaw, K. Hauser, Development of a tele-nursing mobile manipulator for remote care-giving in quarantine areas, in: In: 2017 IEEE International Conference on Robotics and Automation (ICRA), 2017, https://doi.org/10.1109/icra.2017.7989411.
[62] S. Biswas, E. Kruijff, E. Veas, View recommendation for multi-camera demonstration-based training, Multimed. Tools Appl. 83 (2023) 21765–21800, https://doi.org/10.1007/s11042-023-16169-0.
[63] J. Dufek, X. Xiao, R.R. Murphy, Best viewpoints for external robots or sensors assisting other robots, IEEE Trans. Hum. Mach. Syst. 51 (2021) 324–334, https://doi.org/10.1109/thms.2021.3090765.
[64] X. Xiao, J. Dufek, R.R. Murphy, Autonomous visual assistance for robot operations using a tethered UAV, in: Springer Proceedings in Advanced Robotics, Springer Singapore, Singapore, 2021: pp. 15–29. DOI: 10.1007/978-981-15-9460-1_2.
[65] B. Li, B. Lu, Y. Lu, Q. Dou, Y.-H. Liu, Data-driven holistic framework for automated laparoscope optimal view control with learning-based depth perception, in: In: 2021 IEEE International Conference on Robotics and Automation (ICRA), 2021, https://doi.org/10.1109/icra48506.2021.9562083.
[66] S. Yoon, S. Shin, S. Lee, M. Park, C.R. Ahn, Evaluating Viewpoint Control Techniques in Virtual Reality Interface for Teleoperating Construction Welding Robots, in: B. Riveiro, P. Arias (Eds.), Proceedings of the 31st International Workshop on Intelligent Computing in Engineering, European Group for Intelligent Computing in Engineering (EG-ICE), Vigo, Spain, 2024: pp. 345–354. https://3dgeoinfoeg-ice.webs.uvigo.es/proceedings.
[67] K. Kawaharazuka, K. Okada, M. Inaba, Robotic constrained imitation learning for the peg transfer task in Fundamentals of Laparoscopic Surgery, ArXiv [Cs.RO] (2024). http://arxiv.org/abs/2405.03440.
[68] D. Rakita, Methods for Effective Mimicry-based Teleoperation of Robot Arms, in: Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, Association for Computing Machinery, New York, NY, USA, 2017: pp. 371–372. DOI: 10.1145/3029798.3034812.
[69] R. Bellman, A Markovian Decision Process, Indiana Univ. Math. J. 6 (1957) 679–684, https://doi.org/10.1512/IUMJ.1957.6.56038.
[70] C. Celemin, R. Pérez-Dattari, E. Chisari, G. Franzese, L. de Souza Rosa, R. Prakash, Z. Ajanović, M. Ferraz, A. Valada, J. Kober, Interactive imitation learning in robotics: a survey, Found. Tren. Robot. 10 (2022) 1–197, https://doi.org/10.1561/2300000072.
[71] N. Gavenski, F. Meneguzzi, M. Luck, O. Rodrigues, A survey of imitation learning methods, environments and metrics, ArXiv [Cs.LG] (2024). http://arxiv.org/abs/2404.19456.
[72] H. Liu, S. Nasiriany, L. Zhang, Z. Bao, Y. Zhu, Robot learning on the job: Human-in-the-loop autonomy and learning during deployment, in: Robotics: Science and Systems XIX, Robotics: Science and Systems Foundation, 2023. DOI: 10.15607/rss.2023.xix.005.
[73] T. Nguyen, Q. Zheng, A. Grover, Reliable conditioning of behavioral cloning for offline reinforcement learning, ArXiv [Cs.LG] (2022). http://arxiv.org/abs/2210.05158.
[74] Z. Peng, C. Han, Y. Liu, Z. Zhou, Weighted policy constraints for offline reinforcement learning, Proc. Conf. AAAI Artif. Intell. 37 (2023) 9435–9443, https://doi.org/10.1609/aaai.v37i8.26130.
[75] Z. Zhang, Z. Zhuang, J. Xu, D. Wang, M. Liu, S. Zhang, ADR-BC: Adversarial density weighted regression behavior cloning, ArXiv [Cs.LG] (2024). http://arxiv.org/abs/2405.20351.
[76] F. Sasaki, R. Yamashina, Behavioral cloning from noisy demonstrations, in: International Conference on Learning Representations, 2020.
[77] A. Mandlekar, D. Xu, R. Martín-Martín, Y. Zhu, L. Fei-Fei, S. Savarese, Human-in-the-Loop Imitation Learning using Remote Teleoperation, ArXiv [Cs.RO] (2020). http://arxiv.org/abs/2012.06733.
[78] A. Ipsita, L. Erickson, Y. Dong, J. Huang, A.K. Bushinski, S. Saradhi, A. M. Villanueva, K.A. Peppler, T.S. Redick, K. Ramani, in: Towards Modeling of Virtual Reality Welding Simulators to Promote Accessible and Scalable Training, in, Association for Computing Machinery, New York, NY, USA, 2022, pp. 1–21, https://doi.org/10.1145/3491102.3517696.
[79] L. Wei, A. Song, H. Hu, J. Mao, Mixed reality-augmented remote welding system with virtual fixtures and autonomous agents, in: Proceedings of the 2024 4th International Conference on Robotics and Control Engineering, ACM, New York, NY, USA, 2024: pp. 131–137. DOI: 10.1145/3674746.3674767.
[80] Y.-P. Su, X.-Q. Chen, T. Zhou, C. Pretty, G. Chase, Mixed Reality-Enhanced Intuitive Teleoperation with Hybrid Virtual Fixtures for Intelligent Robotic Welding, NATO Adv, Sci. Inst. Ser. E Appl. Sci. 11 (2021) 11280, https://doi.org/10.3390/app112311280.
[81] C. Gokmen, D. Ho, M. Khansari, Asking for help: failure prediction in Behavioral Cloning through value approximation, ArXiv [Cs.RO] (2023). http://arxiv.org/abs/2302.04334.
[82] C. Qian, J. Urain, K. Zakka, J. Peters, PianoMime: Learning a generalist, dexterous piano player from internet demonstrations, ArXiv [Cs.CV] (2024). http://arxiv.org/abs/2407.18178.
[83] Z. Xu, Y. Sun, M. Liu, ICurb: Imitation learning-based detection of road curbs using aerial images for autonomous driving, IEEE Robot. Autom. Lett. 6 (2021) 1097–1104, https://doi.org/10.1109/lra.2021.3056344.
[84] N.M.M. Shafiullah, Z.J. Cui, A. Altanzaya, L. Pinto, Behavior Transformers: Cloning k modes with one stone, ArXiv [Cs.LG] (2022). http://arxiv.org/abs/2206.11251.
[85] C. Wang, L. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu, A. Anandkumar, MimicPlay: Long-horizon imitation learning by watching human play, ArXiv [Cs.RO] (2023). https://mimic-play.github.io/assets/MimicPlay.pdf.