2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

SUBTA: A Framework for Supported User-Guided Bimanual Teleoperation in Structured Assembly

Xiao Liu[1], Prakash Baskaran[1], Songpo Li[1], Simon Manschitz[2], Wei Ma[2], Dirk Ruiken[2], and Soshi Iba[1]

***Abstract*— In human-robot collaboration, shared autonomy enhances human performance through precise, intuitive support. Effective robotic assistance requires accurately inferring human intentions and understanding task structures to determine optimal support timing and methods. In this paper, we present SUBTA, a supported teleoperation system for bimanual assembly that couples learned intention estimation, scene-graph task planning, and context-dependent motion assists. We validate our approach through a user study (N=12) comparing standard teleoperation, motion-support only, and SUBTA. Linear mixed-effects analysis revealed that SUBTA significantly outperformed standard teleoperation in position accuracy ($p < 0.001$, $d = 1.18$) and orientation accuracy ($p < 0.001$, $d = 1.75$), while reducing mental demand ($p = 0.002$, $d = 1.34$). Post-experiment ratings indicate clearer, more trustworthy visual feedback and predictable interventions in SUBTA. The results demonstrate that SUBTA greatly improves both effectiveness and user experience in teleoperation.**

I. INTRODUCTION

Teleoperation refers to operating a machine or robot from a distance, often to allow humans to perform tasks remotely or in hazardous environments. It has proven invaluable in domains such as bomb disposal, underwater exploration, and space missions, where keeping human operators out of danger is paramount [1]. In the manufacturing context, robotic teleoperation enables human workers to execute complex and precise tasks (for example, intricate assembly, welding, or maintenance operations) from a safe location [2]. This approach not only minimizes risks to workers in dangerous or confined factory settings, but also leverages human dexterity and decision-making for tasks that are too variable or difficult to fully automate. Manufacturing teleoperation can thus combine the best of human skill and robotic precision, improving safety and flexibility on the factory floor [2], [3]. However, directly manipulating a robotic arm through a standard interface can be challenging due to differences in the robot's kinematics and sensing compared to a human's, often necessitating extensive training and robotics expertise [4]. Moreover, pure teleoperation demands continuous operator attention and a fast, stable communication link, since the human must control every motion in real-time [5]. These factors make generic teleoperation interfaces cumbersome for manufacturing tasks, especially for operators who are not robotics specialists [5].

[1]Authors are with Honda Research Institute USA, San Jose, CA 95134 USA.
[2]Authors are with Honda Research Institute Europe GmbH, 63073 Offenbach am Main, Germany.
Correspondence: songpo_li@honda-ri.com

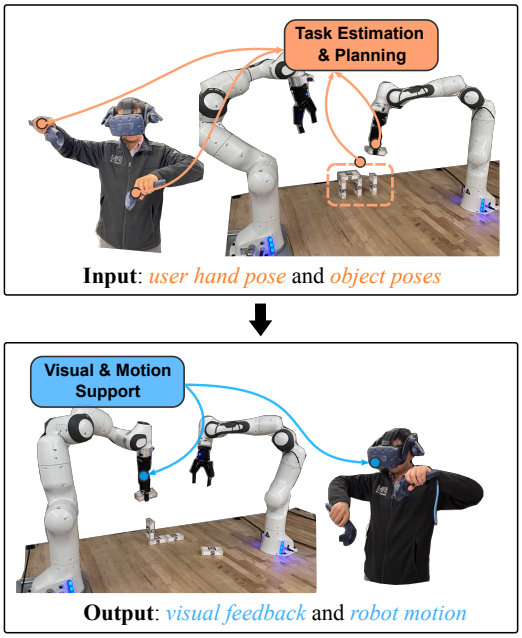


Fig. 1. **SUBTA**: The proposed Supported Teleoperation framework assists users in assembly tasks by (i) estimating the task and generating a plan, and (ii) providing real-time support through visual feedback and motion-level corrections.

To address these challenges, researchers and engineers are turning to task-specific supported teleoperation systems. The central idea is that by supporting the human with intelligent interfaces and autonomy, even non-expert users can perform complex tasks more easily and efficiently than with a one-size-fits-all teleoperation setup [2], [6]. For example, in Figure 1, our system assists a user during a block-assembly task, which is representative of many industrial assembly operations [7], [8]. It estimates user intent, models task state with a scene graph [9], and provides motion support accordingly. Such task-specific support dramatically lowers the barrier to using robots in manufacturing, enabling users to leverage robots for intricate tasks without needing to micromanage every action.

In this paper, we propose a **S**upported **U**ser-Guided **B**imanual **T**eleoperation system for **A**ssembly tasks (SUBTA), extend task-specific support in teleoperation by incorporating task state estimation and graph-based task planning. This enables the robot to understand the task structure in 3D space and assist the user through both visual feedback and motion-level corrections. Our main contributions are summarized as follows:

- Shared autonomy system that integrates three levels of

assistance: task understanding and intention estimation, task planning, and low-level motion support.
- Scene graph representation that encodes spatial relationships in structured assembly tasks for task state estimation and planning.
- Executing a set of gated motion behaviors for grasp and placement, coordinated by a behavior controller to deliver support "when and how" it is needed
- Comprehensive experimental validation with a user study demonstrating the effectiveness of task-specific support in improving teleoperation performance.
- SUBTA reduced cognitive load (6.2 $\rightarrow$ 3.4) while maintaining a 75% success rate, enabling non-experts to perform structured, multi-step assemblies more effectively with nearly twofold higher accuracy.

II. Related Work

Purely manual teleoperation is widely acknowledged to have significant limitations [10]. Operators must continuously control every robot motion, resulting in high cognitive and physical demands, increasing susceptibility to fatigue, and potentially introducing human errors, particularly during prolonged operations [11]. Communication delays or network instabilities exacerbate these issues, as lags in visual feedback or control signals complicate tasks that require precise timing or delicate coordination [12]. Additionally, disparities between the human operator's embodiment and the robotic system—such as differences in scale, degrees of freedom, or sensory inputs—often hinder intuitive control, particularly for complex robotic arms [13].

To address these fundamental limitations, several streams of research have emerged, focusing on improving usability, reducing cognitive load, and enhancing performance in teleoperation systems.

High-level task interfaces represent an early and significant advancement in teleoperation. Rather than relying on low-level, joint-by-joint control, these interfaces allow users to issue commands at a task or semantic level [14]. Researchers have proposed intuitive authoring tools tailored for light manufacturing scenarios, showing that novice users can accomplish complex assembly jobs more efficiently compared to traditional joystick-based methods [15], [16]. These studies consistently indicate that abstracting teleoperation tasks into high-level semantic commands such as "pick-and-place" significantly enhances usability, particularly for users without extensive robotic training.

Shared autonomy has gained traction by introducing real-time robotic assistance guided by intent inference. For instance, Manschitz et al. [4], [17] provided visual and haptic feedback and automated trajectory corrections to help operators accurately perform pick-and-place tasks, especially compensating for common depth-perception errors while maintaining the operator's central role. Similarly, Owan et al. [18] developed dynamic autonomy systems capable of adaptively shifting control between automated execution and human guidance, particularly beneficial in confined-space tasks. These approaches emphasized precision and efficiency while preserving human agency and decision-making capabilities during teleoperation.

Further advancements include *augmented reality* (AR) and *digital twin* technologies, designed to further reduce cognitive load and enhance the operator's situational awareness by visually contextualizing tasks and pre-validating robotic actions. Lu et al. [19] successfully integrated AR headsets with digital twins representing virtual replicas of the robotic work environment. Operators issue high-level commands within this virtual context, which immediately evaluates command feasibility and prevents inappropriate or impossible actions. By providing operators with an informed and interactive task environment, these systems transform robots into collaborative partners that inherently understand operational constraints and offer immediate, actionable feedback [20].

III. Task-specific Teleoperation Framework

The goal of our framework is to support human operators in completing manipulation tasks. In this work, we focus on pick-and-place actions within structured assembly tasks, which require precise object handling to form spatial configurations. The system consists of four main components: 1) Task and Intention Estimation Module: This module takes as input the human operator's motion and the current poses of objects. It estimates the user's intention, including the task being performed and the specific action being taken. 2) Task Planning Module: The current task is modeled as a scene graph to represent the task state accurately. This allows the system to track progress and determine the remaining steps needed to complete the task. 3) Behavior Controller: A state machine that, based on the current context, manages the activation and deactivation of low-level assistance during grasping and placement. It automatically attaches the object to the operator's hand when grasping and snaps it to the target location when the hand is near the desired placement.

A. Task and Intention Estimation

Given SE(3) poses with position and quaternion for orientation of both hands $\mathbf{H}_t \in \mathbb{R}^{6\times 2}$ and n blocks $\mathbf{B}_t \in \mathbb{R}^{6\times n}$, the module predicts the current task label $\mathbf{l}_t$, and left and right-hand actions $\mathbf{a}_t$. The estimation model encodes the human hands $\mathbf{H}_t$, with *tAPE* positional encoding [21], outputs seven 32-dim embeddings (i.e., one embedding per entity) that serve as node features for a graph neural network (GNN). We adapt the structure from the HAR-Transformer [22]. The operator's intention is inferred from evolving spatial relations among entities, modeled as a graph with nodes $\mathbf{V} = \{v_l, v_r, v_{b_1}, \ldots, v_{b_n}\}$, representing the hands and blocks. The adjacency matrix $\mathbf{A}_t$ is dynamic, adapting over time as interactions progress. $\mathbf{A}_t$ is computed using a self-attention mechanism [23] applied to the input node features. Attention scores are scaled, normalized via softmax, and symmetrized by averaging with their transpose:

$$\begin{aligned} &\mathit{Attn}(Q,K,V) = \mathit{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right), \quad \mathit{softmax}(x_i) = \frac{e^{x_i}}{\sum e^{x_i}} \\ &\mathbf{A}_t = (\mathit{Attn} + \mathit{Attn}^T)/2\,, \end{aligned} \tag{1}$$

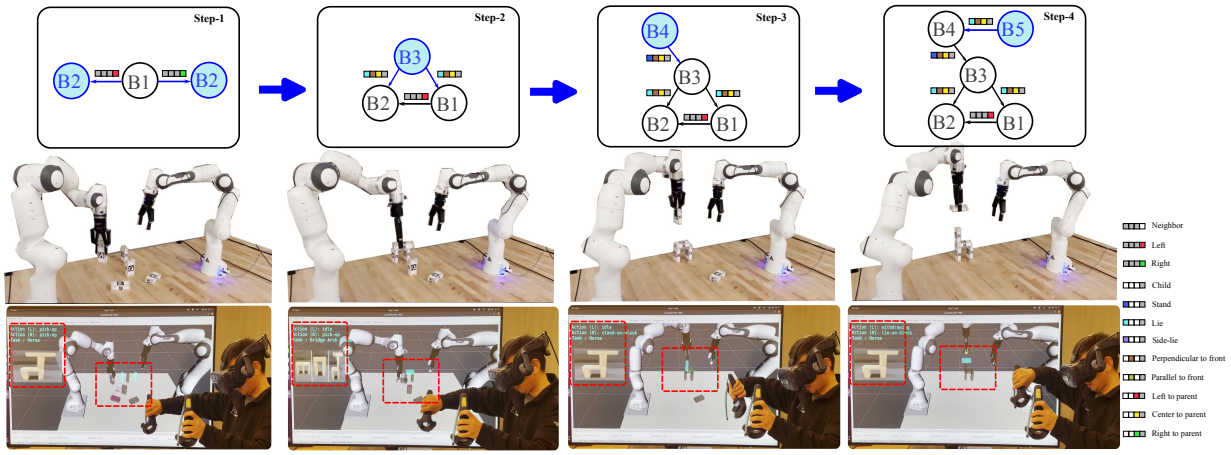


Fig. 2. **Supported teleoperation system**: The user teleoperates the robot to assemble the "horse" structure. The **top row** shows the scene graph encoding spatial relationships for task monitoring and planning. The **middle row** depicts the robot successfully executing the block assembly. The **bottom row** presents the digital twin environment, where task estimation and planning ((highlighted in red) are visualized to guide the user.

where, Q, K, V are the input features and d_k their dimensionality. The resulting $\mathbf{A}_t$ captures learned, symmetric spatial relationships and is optimized end-to-end. The GNN extracts spatial features by applying graph convolutions across K layers, each followed by an activation (except the final layer). The layer-wise computation is:

$$\begin{aligned} \boldsymbol{Z}_t^{(k)} &= f_k\left(\mathbf{A}_t, \boldsymbol{F}_t^{(k-1)}\right), \\ \boldsymbol{Z}_t^{(k)} &= \mathbf{A}_t * \boldsymbol{F}_t^{(k-1)} * \boldsymbol{W}^{(k-1)} \\ \boldsymbol{F}_t^{(k)} &= \alpha_k\left(\boldsymbol{Z}_t^{(k)}\right), \end{aligned} \tag{2}$$

where $\mathbf{A}_t \in \mathbb{R}^{(n+2)\times(n+2)}$ is the adjacency matrix, and $\boldsymbol{F}^{(0)}$ is the GNN's input node feature layer. Each layer outputs $\boldsymbol{F}_t^{(k)} \in \mathbb{R}^{(n+2)\times d_k}$, with learnable weights $\boldsymbol{W}^{(k-1)} \in \mathbb{R}^{d_{k-1}\times d_k}$. The final GNN output $\mathbf{F}^{(2)}$ is flattened, passed through a 32-unit ReLU layer, and forms a feature vector $\mathbf{g}$. This vector feeds three heads that predict (i) the task (sigmoid), (ii) left-hand action, and (iii) right-hand action. The action heads concatenate $\mathbf{g}$ with velocity embeddings from a three-layer HAR-Transformer [22]. The training is conducted with Adam (lr = 10^{-3}), using class-weighted binary cross-entropy for the task head and focal loss for the action heads to address class imbalance. Inputs are 3 sec windows (20 Hz) with a 1 sec stride.

B. Task Planning

Scene graphs have been shown to effectively capture diverse object relationships in manipulation tasks [24]. In this work, we extend this idea by employing scene graphs to represent the geometric structure of assembly tasks. Given $\mathbf{B}_t \in \mathbb{R}^{7\times n}$ as the set of SE(3) poses used in one assembly. We design a heuristic, rule-based algorithm h to extract the pairwise spatial relationships between blocks. When constructing the scene graph, we create edges only between blocks that are in contact or at the same height. We represent the scene graph as:

$$\begin{aligned} &\mathcal{G}_t = (\boldsymbol{V}_t, \boldsymbol{E}_t, \boldsymbol{\epsilon}_t, h), \\ &h : \boldsymbol{\epsilon}_t \rightarrow \{\{v_n, v_m\} | v_n, v_m \in \boldsymbol{V}_t \text{ and } h(v_n, v_m) \neq \text{None}\}. \end{aligned} \tag{3}$$

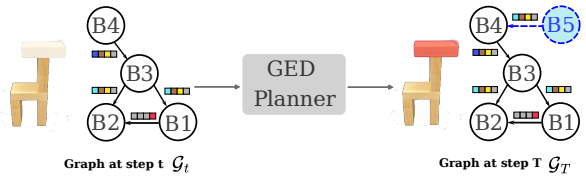


Fig. 3. **Task Planner**: the planner computes the most likely target pose for the next block based on the graph edit distance (GED) and adds the corresponding node to the graph.

TABLE I

DISCRETE SPATIAL-RELATION EDGE ATTRIBUTES.

Label	Ori. w.r.t. Parent	Ori. w.r.t. Front	Pos. w.r.t. Parent	Pos. w.r.t. Neighbor
0	–	–	–	–
1	Stand	Parallel	Center	Left
2	Lie	Perpendicular	Left	Right
3	Side-lie	–	Right	–

Ori. = Orientation, Pos. = Position

Each node $v_t \in \boldsymbol{V}_t$ carries the raw pose of its block. When the operator begins manipulating a new block, a new node for that block is appended to the graph, and all incident edges $e_t \in \boldsymbol{E}_t$ together with its attributes $\epsilon_t \in \boldsymbol{\epsilon}_t$ are generated by the heuristic h before being added. Edge attributes, defined in Table I, encode the spatial relationships between blocks during assembly. For example, in Figure 3, block B_5 and block B_4 are connected with the attribute $\epsilon_t = [2, 2, 1, 0]$, indicating that B_5 is placed in lying pose and centrally on top of B_4 and oriented perpendicular to its front face.

The task planning is achieved by using graph edit distance (GED). The planning and next step target prediction can be combined to automatically complete the task at any stage of a trial. We define a set of graph edit operations $\pi = \{o_1, o_2, o_3\}$, where o_1 refers to add-delete-modify any node v_i, o_2 refers to add-delete-modify any edge e_i, and o_3 refers to add-delete-modify any edge attributes ϵ_i. Given a current scene graph $\mathcal{G}_t$, and the end scene graph $\mathcal{G}_T$, we use the algorithm

$$d_{\text{GED}}(\mathcal{G}_t, \mathcal{G}_T) = \min_{\pi \in \mathcal{P}(\mathcal{G}_t, \mathcal{G}_T)} \sum_{o \in \pi} c(o), \tag{4}$$

TABLE II

CONTEXT-DEPENDENT BEHAVIOURS FOR PICK–PLACE MOTION SUPPORT.

#	Behaviour	Trigger	User control	System autonomy / constraints	Feedback
1	Approach Object	Intention to pick & object detected	6-DoF free motion	Limit joint/speed for safety;	-
2	Snap to Object	Hand to object dist. $< \delta_1$	Frozen	Auto-drive EE to grasp pose	Object highlight
3	Align w/ Object	Snap to object done	Nullspace motion	Collision avoidance in grasp manifold	-
4	Grasp Object	Button press	EE locked	Close gripper	Haptic click
5	Align w/ Surface	Object grasped, on plane	on-plane transformation	Keep contact; suppress lift/tilt	Plane highlight
6	Unsnap Surface	Controller lifted $> \delta_2$	-	Safe lift motion	-
7	Approach Surface	Unsnap surface done	6-DoF free motion	Limit joint/speed for safety	-
8	Snap to Surface	Hand to plane dist. $< \delta_3$	Frozen	Auto-drive EE to plane, Keep contact	Plane highlight
9	Release Object	Finger-opend	EE locked	Open fingers; release	-

End-effector (EE), distance (dist.), δ_1, δ_2 and δ_3 are adjustable thresholds.

where $\mathcal{P}(\mathcal{G}_t, \mathcal{G}_T)$ is the set of edit paths that transform $\mathcal{G}_t$ into $\mathcal{G}_T$, $c(o) \geq 0$ is the cost assigned to operation o. Figure 3 illustrates the procedure for the "horse" assembly task. At each time step the planner constructs the current scene graph $\mathcal{G}_t$, compares it with the goal graph $\mathcal{G}_T$, and produces both the graph-edit distance d_{GED} and the next intermediate graph on the optimal edit path. The goal graph is selected by the task estimation module, where one assembly can have multiple goal graphs. In the right panel of Figure 3, block B_5 (highlighted in blue) is selected as the next target. After a target node v_i is selected, the algorithm retrieves the corresponding parent and adjacent nodes using simply Breadth-first search. A rule-based heuristic then estimates the required pose for v_i based on the edge attributes, and renders this pose in the digital-twin environment, providing real-time guidance to the operator.

C. Motion Support Behaviors

We decompose the overall pick-and-place routine into nine context-dependent behaviors. Transitions between behaviors are governed by inferred operator intention, hand–object distance, and motion feasibility. This modular design enables graduated support: full manual control is retained during coarse motions, while autonomous corrections are introduced in precision phases. Details of each behavior are summarized in Table II. Visual feedback to the operator includes highlighted objects during snapping, emphasized block surfaces for alignment, and snapping indicators to surfaces during placement.

IV. EXPERIMENTS DESIGN

A. Experimental setup

Manipulation sequences were collected from humans performing block assembly tasks through teleoperation of a bimanual robotic system in a virtual reality environment [4], [25]. Following the setup described in [26], participants executed eight distinct block assembly tasks within a virtual scene rendered in Rviz and displayed using an HTC Vive Pro Eye headset. The scene consisted of a table with five identical wooden blocks available for assembly. Block poses were reliably tracked using ArUco markers detected by three Intel RealSense D435 cameras.

Dataset for task estimation module was collected from 495 teleoperation demonstrations form various block assembly

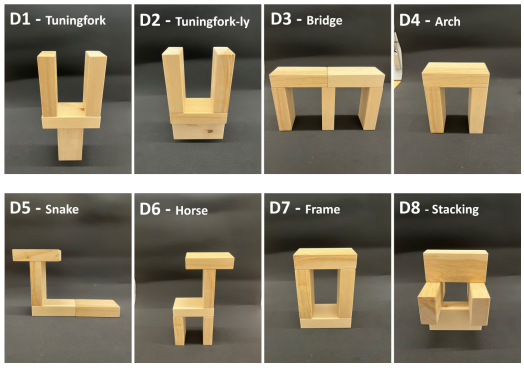


Fig. 4. Eight Block assembly tasks.

TABLE III

COMPARISON OF F1-SCORES (% MEAN $\pm$ STD. DEV.)

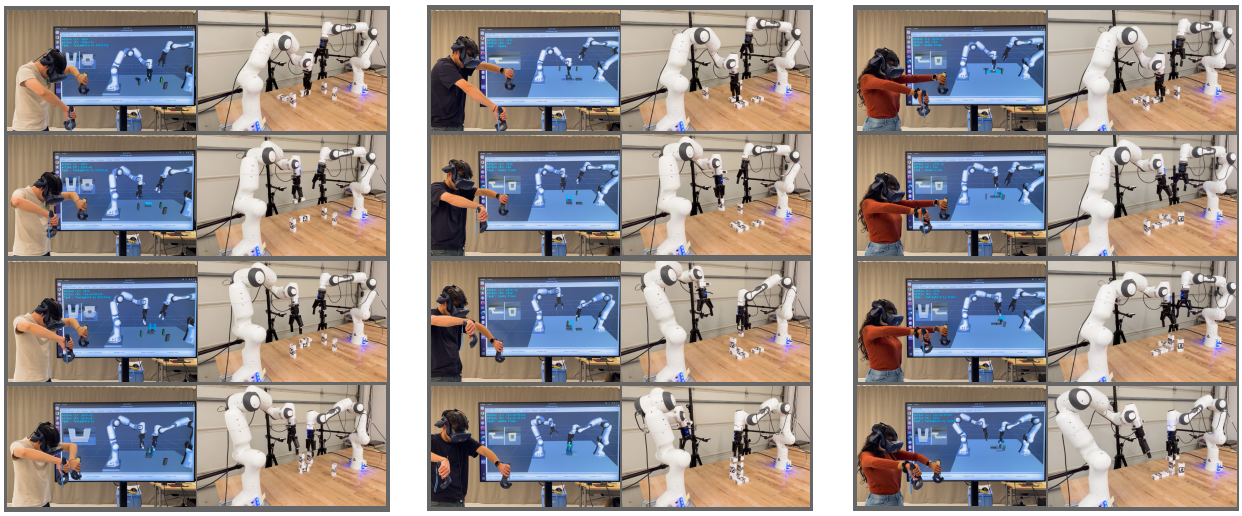
Model	Action (L)	Action (R)	Task
Naive CNN	73.53 (6.40)	72.34 (5.61)	77.58 (4.38)
Hierarchical InEs	43.97 (10.97)	38.93 (8.90)	49.43 (8.62)
HAR-Transformer	81.49 (4.55)	79.95 (4.39)	**90.12 (3.71)**
Ours	**81.73 (4.47)**	**80.48 (4.84)**	89.82 (3.75)

tasks. Nine distinct actions defined by end-effector movements were performed to complete these tasks: Idle, Pick-up, Withdraw, Stand, Lie, Side-lie, Stand-on-block (Stand-OB), Lie-on-block (Lie-OB), and Side-lie-on-block (Side-lie-OB). Each teleoperated demonstration required executing an entire task through action sequences selected freely by the participants. The initial positions and orientations of the blocks were randomized prior to each demonstration. Left and right actions were labeled independently according to their start and end times, while task labels were multi-labeled due to partial or complete overlaps, such as the Arch task overlapping with Bridge and Horse tasks (see Figure 4). The dataset consists of totaling 480.4 minutes, and includes recordings of SE(3) poses of five wooden blocks, end-effector poses, and egocentric videos of teleoperators.

B. Task Estimation Evaluation

1) Baselines: The proposed GNN networks are benchmarked against three baselines: i) Naive CNN architecture, a three-layered 1-D convolutional network, ii) *Hierarchical InEs* [27], and iii) State-of-the-art *HAR-Transformer* [22]. The latter two models were adapted appropriately.

2) Evaluation: The intention algorithms were validated using the *leave-one-subject-out* cross-validation scheme,



(a) "Tuningfork-ly" task **(b)** "Frame" task **(b)** "Snake" task

Fig. 5. Examples of participants performing three different assembly tasks using the SUBTA system during the user study. The supported workspace is projected on the left with the user included, while the actual robot workspace is shown on the right.

where the mean F1-score is reported by training the algorithms repeatedly on all, but one participant's data and validated using the left-out participant's data [28].

Table III shows that the proposed model substantially outperforms the naive CNN and hierarchical baselines, achieving performance on par with or exceeding the state-of-the-art HAR-Transformer for task and action prediction. Across models, left-hand predictions are 2–4% more accurate than right-hand predictions, reflecting the predominance of right-handed participants whose left-hand activity is skewed toward the easier-to-predict Idle class. Finally, task predictions consistently surpass action predictions, indicating that short-term action prediction is inherently more difficult due to its transient and dynamic nature. We then integrate this module into the SUBTA system for more comprehensive user studies.

C. Subject Study

1) The goal of the study: The goal of this study is to evaluate the accuracy and effectiveness of the proposed SUBTA system against standard teleoperation in robotic assembly tasks. We compare three modes:

1) **Standard teleoperation** (M1): direct retargeting with no additional support.
2) **Motion-support only** (M2): an ablation of SUBTA providing only general motion assistance.
3) **SUBTA** (M3): with both task-specific and motion support.

System performance is evaluated based on task success rate, completion time, and target pose accuracy. In addition to these objective metrics, participants provide subjective feedback through the NASA Task Load Index (NASA-TLX) and the System Usability Scale (SUS) [29]. After completing all modes, participants also complete another survey rating their overall experience with each interface.

2) Experimental Hypothesis: We formulate the following primary hypotheses for our study: ($\boldsymbol{H_1}$) the proposed system increases the quality of the assembly for participants; ($\boldsymbol{H_2}$) the proposed system achieves higher success rates in robotic assembly tasks compared to standard teleoperation methods; ($\boldsymbol{H_3}$) the proposed system is intuitive and imposes a lower cognitive workload on users during task execution; and ($\boldsymbol{H_4}$) the proposed system show improved usability compare to standard teleoperation methods.

TABLE IV

LINEAR MIXED-EFFECTS MODEL RESULTS ACROSS THE THREE MODES

Hypothesis	M3 vs. M2	M2 vs. M1	M3 vs. M1	Supported
$\boldsymbol{H_1a}$: Pos. err.	*	**	**	Yes
$\boldsymbol{H_1b}$: Ori. err.	**	**	**	Yes
$\boldsymbol{H_3}$: Mental	NS	**	**	Yes
$\boldsymbol{H_4}$: SUS	NS	*	*	Yes

NS (not significant); ** $p < 0.01$; * $0.01 < p < 0.05$.

3) Participant Study: We evaluated SUBTA (M3) in comparison with the motion-support mode (M2) and standard teleoperation (M1) under the experimental setup described in Section IV-A. All procedures complied with relevant ethical guidelines and regulations. Participants were eligible if they were (1) at least 18 years old, (2) free of known physical or cognitive impairments, and (3) had no history of musculoskeletal disorders. Examples from the study are shown in Figure 5. Prior to testing, each participant completed a practice session with all modes. During the experiment, participants performed three randomly selected tasks out of four under each mode. The order of modes was randomized, and participants advanced once they reported being comfortable.

We recruited $N = 12$ adult participants. 12 participants (11 male, 1 female) took part in the study, with ages ranging from 24 to 44 years (mean 33). On average, participants reported low gaming frequency (2.8/7) and moderate experience with robot teleoperation (3.7/7). Based on pilot/observed effect sizes, an a priori power analysis for our planned *pairwise* contrasts against the baseline mode (M1) indicated that, with two-sided $\alpha = .05$, approximately 7–11 participants would achieve 80% power for the subjective

TABLE V

PER-TASK AND OVERALL SUMMARIES (MEAN ± STD) BY MODE. BEST VALUE PER TASK IS BOLDED.

Task	Mode	Time (sec)	Success Rate	Progress to Complete	Orientation Error (deg)	Position Error (meter)
"Tuningfork-ly"	M1	74.56 ± –	11.1%	38.9% ± 25.3%	3.710 ± –	0.022 ± –
	M2	**71.02** ± –	12.5%	43.8% ± 29.1%	2.540 ± –	0.013 ± –
	M3	87.05 ± 38.26	**22.2%**	**50.0% ± 33.1%**	**0.920 ± 0.382**	**0.011 ± 0.006**
"Arch"	M1	64.14 ± 13.51	62.5%	74.9% ± 34.7%	5.288 ± 3.310	0.028 ± 0.018
	M2	71.32 ± 15.13	85.7%	95.1% ± 12.9%	2.568 ± 2.032	0.013 ± 0.004
	M3	**57.86 ± 7.17**	**100.0%**	**100.0% ± 0.0%**	**2.104 ± 2.224**	**0.007 ± 0.004**
"Snake"	M1	95.03 ± 20.01	60.0%	87.5% ± 17.7%	4.272 ± 2.599	0.038 ± 0.030
	M2	**83.19 ± 20.89**	**88.9%**	**97.2% ± 8.3%**	**1.459 ± 0.816**	0.027 ± 0.013
	M3	88.16 ± 31.59	80.0%	92.5% ± 16.9%	1.836 ± 1.103	**0.023 ± 0.007**
"Frame"	M1	88.61 ± 15.61	88.9%	91.7% ± 25.0%	5.394 ± 2.572	0.034 ± 0.019
	M2	**79.98 ± 14.42**	88.9%	94.4% ± 16.7%	3.194 ± 2.333	0.023 ± 0.016
	M3	88.39 ± 22.93	**100.0%**	**100.0% ± 0.0%**	**2.073 ± 1.641**	**0.019 ± 0.015**
Overall	M1	83.71 ± 19.73	55.6%	73.6% ± 32.6%	4.947 ± 2.619	0.033 ± 0.022
	M2	**78.45 ± 16.81**	69.7%	83.1% ± 28.5%	2.399 ± 1.855	0.021 ± 0.013
	M3	79.18 ± 26.45	**75.0%**	**85.4% ± 27.6%**	**1.927 ± 1.610**	**0.016 ± 0.012**

outcomes (e.g., TLX Mental, SUS total) [30]. In addition, the objective *accuracy* outcomes (position/orientation) error exhibited large effects when compared to M1, for which our within-subject design provides high sensitivity; thus our final sample of $N = 12$ was adequate for these primary comparisons.

V. RESULTS AND DISCUSSION

In what follows, we compare the proposed SUBTA system against two baseline modes using both objective and subjective outcomes, and evaluate how these results bear on the hypotheses in Section IV-C.2. The definition and computation of each metric are provided in the following sections. Linear mixed-effects model results are summarized in Table IV; note that the pose accuracy is evaluated using position errors and orientation errors, shown in the first and second rows. Overall, SUBTA (M3) outperforms standard teleoperation (M1): it achieves significantly higher assembly accuracy, lower reported mental workload (TLX–Mental), and higher SUS scores. Therefore, $\boldsymbol{H_1}$, $\boldsymbol{H_3}$, and $\boldsymbol{H_4}$ are supported. For $\boldsymbol{H_2}$, although the M3 vs. M1 comparison in success rate emphasizes practical rather than statistical significance, M3 achieves a substantial absolute gain in completion (55.6% $\rightarrow$ 75%), so we regard $\boldsymbol{H_2}$ as supported as well.

A. Quantitative Analysis

The quantitative results are summarized in Table V. Task progress is measured as the number of correctly placed blocks relative to the total blocks in the assembly. We treat the first placed anchor block as defining the global reference frame for that trial; all subsequent target poses are computed by applying the known relative transforms from $\mathcal{G}_T$ to that anchor. Position and orientation errors are computed by comparing user-placed block poses against the ground-truth poses. Linear mixed-effects analysis revealed that SUBTA (M3) significantly outperformed standard teleoperation (M1) in position accuracy. A linear mixed-effects model revealed significant improvements in position accuracy ($t(99) = 7.88$, $p < 0.001$, $d = 1.18$) and orientation accuracy ($t(99) = 12.67$, $p < 0.001$, $d = 1.75$), as well as a significant reduction in mental demand ($t(99) = 4.65$, $p = 0.002$, $d = 1.34$. M3 delivered up to a two-fold gain in pose accuracy relative to M1. Notably, M2 provides advantages in specific tasks, for example, in the "snake" assembly, participants achieved 97.2% completion progress. M2 also yields slightly faster overall completion times (79.18 sec $\rightarrow$ 78.45 sec) than M3, suggesting that motion-level assistance can facilitate quicker grasping and placement compared to baseline teleoperation in general.

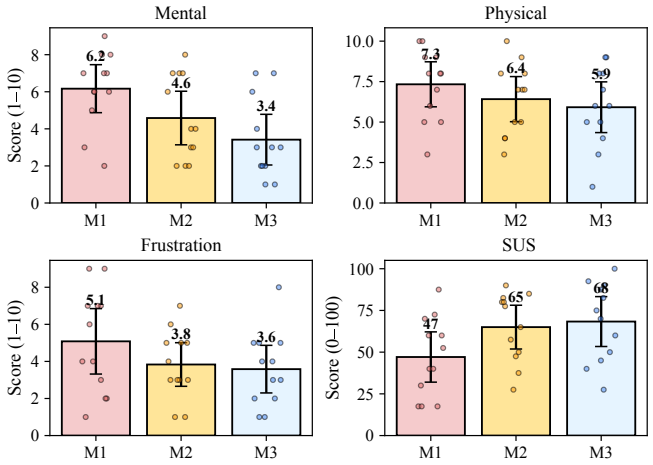


Fig. 6. NASA-TLX ratings for mental demand, physical demand, and frustration, along with SUS scores across the three modes.

B. Survey Responses

As shown in Figure 6, we summarize subjective workload and usability across the three modes. TLX dimensions (Mental, Physical, Frustration) are scored on 1–10 (lower is better), and SUS on 0–100 (higher is better). We computed the sample mean $\bar{x}$ and standard deviation s and formed a 95% confidence interval (CI) for the population mean using a t-distribution: $\bar{x} \pm t_{0.975,\,n-1}\,\frac{s}{\sqrt{n}}$, where n is the sample size. These intervals are shown as the vertical error bars in

TABLE VI

POST-EXPERIMENT QUESTIONNAIRE ITEMS FOR M2 AND M3 ON A 10-POINT LIKERT SCALE (1 = STRONGLY DISAGREE, 10 = STRONGLY AGREE).

Category	M2 Items	M3 Items
Visual	(1) Placement surface visualization reduced ambiguity (2) Placement surface visualization increased my confidence	(1) Predicted target block visualization reduced ambiguity (2) Predicted target block visualization increased my confidence (3) Visualization was accurate and trustworthy (4) Visualization was uncluttered
Motion	(3) Hand snapping made grasping easier (4) Object snapping made placement easier (5) Snapping increased my confidence (6) Interventions were predictable and transparent	Same as M2 (items 3–6)
Agency	(7) I always felt in control (8) Easy to override/disengage snapping	Same as M2 (items 7–8)

the figure. The jittered points indicate the responses from individual participants.

Mental demand. We observe the means decrease monotonically (M1 $>$ M2 $>$ M3), with visibly tighter CIs for M3. Paired-samples comparisons relative to M1 indicate substantial reductions in perceived cognitive effort: M3 vs. M1, $p<0.01$; M2 vs. M1, $p<0.01$; M3 vs. M2, not significant ($p>0.05$). These results suggest that SUBTA markedly lowers mental workload compared to standard teleoperation.

Physical demand. Significant differences were also observed for *physical demand* across the three modes. The linear mixed-effects model result shows M3 vs. M1: $p<0.01$ and M2 vs. M1: $p \approx 0.05$. Participant-level points cluster at lower values under M3, indicating more consistently reduced effort. Overall, both M2 and M3 reduce physical demand, with M3 providing the greatest reduction.

Frustration. The same pattern holds for frustration: M3 has the lowest mean and a tighter distribution, indicating less irritation or stress than the standard teleoperation, while M2 remains intermediate between M1 and M3.

Usability. SUS (0–100) measures perceived usability. The mean scores rise monotonically (M1 $<$ M2 $<$ M3) from all participants. The linear mixed-effects model result shows the same pattern: M3 vs. M1, $p \approx .001$; M2 vs. M1, $p \approx .001$; M3 vs. M2, not significant ($p>0.05$). Thus, both assisted modes are rated more usable than standard teleoperation mode, and M3 has the highest mean, though not reliably higher than M2. For context, a SUS of about 68 is commonly treated as "average" usability [29]; SUBTA (M3) exceeds this benchmark and clearly improves over standard teleoperation (M1).

Overall, participants report *lower* mental/physical workload and frustration and *higher* usability in SUBTA (M3) compared to standard teleoperation (M1), with motion-support (M2) providing intermediate benefits.

C. Visual, Motion, and Agency Feedback

The post-experiment questions are summarized in Table VI. We analyzed post-experiment ratings for the M2 and M3 across three domains, *Visual*, *Motion*, and *Agency*. The distributions of the responses are visualized in Figure 7.

Visual feedback: M3 provides substantially stronger visual support than M2. On the composite visual score, the

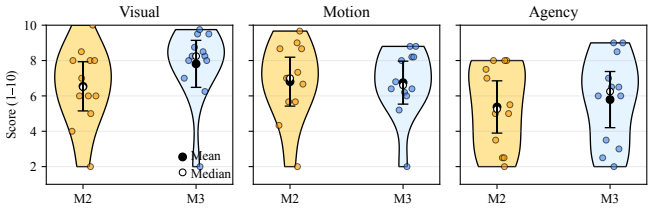


Fig. 7. Full distributions with mean, median, and 95% confidence intervals for the post-experiment comparison of M2 and M3.

mean increased from $\approx$ 6.5 in M2 to $\approx$ 7.8 in M3 (shown in the figure), with a medium within-subject effect (paired trend: $p \approx .08$). Participants thus rated the *predicted-target-block* visualization as clearer and more helpful than the *placement-surface* visualization in M2. This aligns with our broader results: M3's visual guidance reduces ambiguity about what to act on and boosts confidence in teleoperation. Practically, these findings argue for predictive visual cues that are accurate, uncluttered, and specific to the intended target.

Motion feedback: Composite ratings for snapping behaviors and intervention predictability were comparable in M2 and M3 (means $\approx$ 6.8 vs. $\approx$ 6.8, $p \approx .89$). This suggests that the *motor-level assistance* (e.g., hand/object snapping, placement assistance, and transparency of interventions) was already effective in M2 and did not markedly change in perceived helpfulness when moving to M3. A design implication is that further gains in motion-support feedback may require more adaptive triggering when a snap occurs.

Agency feedback: Agency ratings ("felt in control" and "easy to override/disengage") show a small improvement from M2 to M3 (means $\approx 5.4 \rightarrow 5.8$, paired trend $p \approx .12$). Although not statistically significant with $N = 12$, the shift and the violin distributions indicate that users did not feel overruled by M3's additional assistance, instead, the users reported M3 feels *slightly greater* in control. Together with the predictability items, this supports the view that M3's interventions were understandable and that override remained accessible.

Taken together, these subjective data indicate that M3's main added value comes from its *visual* guidance: clearer target specification and improved user confidence. Perceived *motion* assistance is already strong in M2 and remains com-

parable in M3, while *agency* shows a modest upward shift in favor of M3. In combination with previous objective results, the evidence supports **SUBTA (M3) as the preferable mode: it leverages predictive visual feedback without eroding user control or transparency**.

VI. Conclusion

We presented SUBTA, a supported teleoperation framework that integrates learned task and intention estimation, scene-graph–based task planning, and context-dependent motion support for bimanual assembly. In the user study, SUBTA achieved significantly higher pose accuracy (position and orientation), lower NASA–TLX mental load, and higher usability scores (SUS) compared to standard teleoperation. It also improved success rates in absolute terms (55.6% $\rightarrow$ 75%). The study further demonstrated the benefits of systematically combining visual and motion support, showing that clear target specification improved user confidence. Future work will include larger-scale user studies and the development of adaptive policies to enhance perceived motion support and user agency while maintaining system predictability.

References

[1] Jianhong Cui, Sabri Tosunoglu, Rodney Roberts, Carl Moore, and Daniel W Repperger. A review of teleoperation system control. In *Proceedings of the Florida conference on recent advances in robotics*, pages 1–12. Citeseer, 2003.

[2] Chen Zheng, Kangning Wang, Shiqi Gao, Yang Yu, Zhanxi Wang, and Yunlong Tang. Design of multi-modal feedback channel of human–robot cognitive interface for teleoperation in manufacturing. *Journal of Intelligent Manufacturing*, pages 1–21, 2024.

[3] Claudia González, J Ernesto Solanes, Adolfo Munoz, Luis Gracia, Vicent Girbés-Juan, and Josep Tornero. Advanced teleoperation and control system for industrial robots based on augmented virtuality and haptic feedback. *Journal of Manufacturing Systems*, 59:283–298, 2021.

[4] Simon Manschitz and Dirk Ruiken. Shared autonomy for intuitive teleoperation. In *ICRA Workshop: Shared Autonomy in Physical Human-Robot Interaction: Adaptability and Trust*, 2022.

[5] Emmanuel Senft, Michael Hagenow, Kevin Welsh, Robert Radwin, Michael Zinn, Michael Gleicher, and Bilge Mutlu. Task-level authoring for remote robot teleoperation. *Frontiers in Robotics and AI*, 8:707149, 2021.

[6] Emmanuel Akita, Guy Zaidner, and Mitch Pryor. Improved situational awareness and performance with dynamic task-based overlays for teleoperation. In *Proceedings of the 2024 International Symposium on Technological Advances in Human-Robot Interaction*, pages 65–73, 2024.

[7] Manav Kulshrestha and Ahmed H Qureshi. Structural concept learning via graph attention for multi-level rearrangement planning. In *Conference on Robot Learning*, pages 3180–3193. PMLR, 2023.

[8] Takuya Kiyokawa, Naoki Shirakura, Zhenting Wang, Natsuki Yamanobe, Ixchel G Ramirez-Alpizar, Weiwei Wan, and Kensuke Harada. Difficulty and complexity definitions for assembly task allocation and assignment in human–robot collaborations: A review. *Robotics and Computer-Integrated Manufacturing*, 84:102598, 2023.

[9] Hanxiao Jiang, Binghao Huang, Ruihai Wu, Zhuoran Li, Shubham Garg, Hooshang Nayyeri, Shenlong Wang, and Yunzhu Li. Roboexp: Action-conditioned scene graph via interactive exploration for robotic manipulation. *arXiv preprint arXiv:2402.15487*, 2024.

[10] Daniel J Rea and Stela H Seo. Still not solved: A call for renewed focus on user-centered teleoperation interfaces. *Frontiers in Robotics and AI*, 9:704225, 2022.

[11] Erwin Jose Lopez Pulgarin, Ozan Tokatli, Guy Burroughes, and Guido Herrmann. Assessing tele-manipulation systems using task performance for glovebox operations. *Frontiers in Robotics and AI*, 9:932538, 2022.

[12] Parinaz Farajiparvar, Hao Ying, and Abhilash Pandya. A brief survey of telerobotic time delay mitigation. *Frontiers in Robotics and AI*, 7:578805, 2020.

[13] Anca D Dragan, Siddhartha S Srinivasa, and Kenton CT Lee. Teleoperation with intelligent and customizable interfaces. *Journal of Human-Robot Interaction*, 2(2):33–57, 2013.

[14] Lingxiao Meng, Jiangshan Liu, Wei Chai, Jiankun Wang, and Max Q-H Meng. Virtual reality based robot teleoperation via human-scene interaction. *Procedia Computer Science*, 226:141–148, 2023.

[15] Portia Wang, Shreeya Jain, Manxueying Li, Shutaro Aoyama, Xuezhen Wang, Shuran Song, Jen-Shuo Liu, and Steven Feiner. Built to order: A virtual reality interface for assigning high-level assembly goals to remote robots. In *Proceedings of the 2023 ACM Symposium on Spatial User Interaction*, pages 1–2, 2023.

[16] Shutaro Aoyama, Jen-Shuo Liu, Portia Wang, Shreeya Jain, Xuezhen Wang, Jingxi Xu, Shuran Song, Barbara Tversky, and Steven Feiner. Asynchronously assigning, monitoring, and managing assembly goals in virtual reality for high-level robot teleoperation. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pages 450–460. IEEE, 2024.

[17] Simon Manschitz, Berk Gueler, Wei Ma, and Dirk Ruiken. Sampling-based grasp and collision prediction for assisted teleoperation. In *IEEE International Conference on Robotics and Automation*, 2025.

[18] M. B. Owan et al. Dynamic autonomy for shared human-robot control in confined spaces. *IEEE Robotics and Automation Letters*, 2022.

[19] Yuqian Lu, Chao Liu, I Kevin, Kai Wang, Huiyue Huang, and Xun Xu. Digital twin-driven smart manufacturing: Connotation, reference model, applications and research issues. *Robotics and computer-integrated manufacturing*, 61:101837, 2020.

[20] Zequn Zhang, Yuchen Ji, Dunbing Tang, Jie Chen, and Changchun Liu. Enabling collaborative assembly between humans and robots using a digital twin system. *Robotics and Computer-Integrated Manufacturing*, 86:102691, 2024.

[21] Navid Mohammadi Foumani, Chang Wei Tan, Geoffrey I Webb, and Mahsa Salehi. Improving position encoding of transformers for multivariate time series classification. *Data Mining and Knowledge Discovery*, 38(1):22–48, 2024.

[22] Iveta Dirgová Luptáková, Martin Kubovčík, and Jiří Pospíchal. Wearable sensor-based human activity recognition with transformer model. *Sensors*, 22(5):1911, 2022.

[23] A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A Gomez, L Kaiser, and I Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[24] Guangyao Zhai, Xiaoni Cai, Dianye Huang, Yan Di, Fabian Manhardt, Federico Tombari, Nassir Navab, and Benjamin Busam. Sg-bot: Object rearrangement via coarse-to-fine robotic imagination on scene graphs. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4303–4310. IEEE, 2024.

[25] Anna Belardinelli, Anirudh Reddy Kondapally, Dirk Ruiken, Daniel Tanneberg, and Tomoki Watabe. Intention estimation from gaze and motion features for human-robot shared-control object manipulation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 9806–9813. IEEE, 2022.

[26] Prakash Baskaran, Xiao Liu, Songpo Li, and Soshi Iba. explainable intention estimation in teleoperated manipulation using deep dynamic graph neural networks. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 16551–16558. IEEE, 2025.

[27] Mingyu Cai, Karankumar Patel, Soshi Iba, and Songpo Li. Hierarchical deep learning for intention estimation of teleoperation manipulation in assembly tasks. In *IEEE International Conference on Robotics and Automation*, pages 17814–17820. IEEE, 2024.

[28] Trevor Hastie. The elements of statistical learning: data mining, inference, and prediction, 2009.

[29] James R Lewis. The system usability scale: past, present, and future. *International Journal of Human–Computer Interaction*, 34(7):577–590, 2018.

[30] R Barker Bausell and Yu-Fang Li. *Power analysis for experimental research: a practical guide for the biological, medical and social sciences*. Cambridge University Press, 2002.