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

Adaptor: Advancing Assistive Teleoperation with Few-Shot Learning and Cross-Operator Generalization

Yu Liu[1], Yihang Yin[2], Tianlv Huang[1], Fei Yan[1], Yuan Xu[1], Weinan Hong[1], Wei Han[1], Yue Cao[2], Xiangyu Chen[2], Zipei Fan[1,†], and Xuan Song[1]

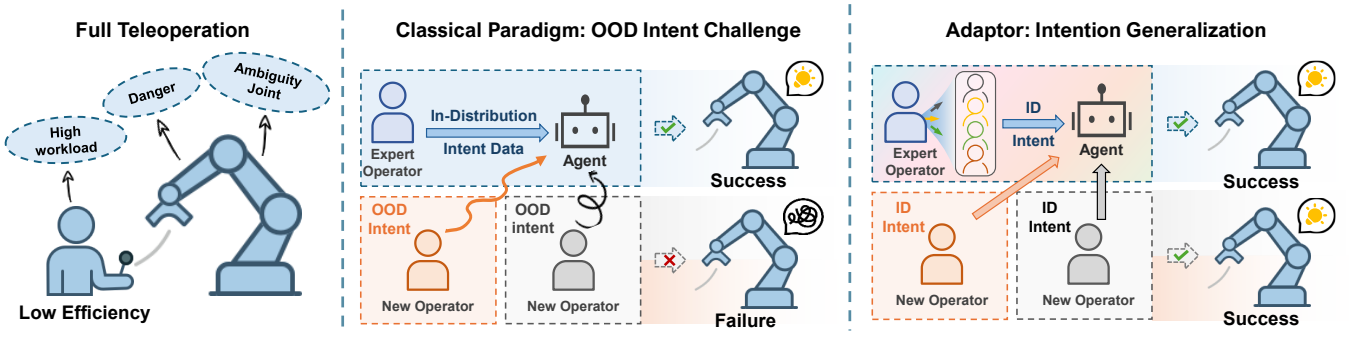

Fig. 1. **Evolution of teleoperation paradigms.** Left: Direct teleoperation maps human inputs to robot commands but suffers from instability due to human-robot dynamic mismatches. Middle: Conventional assistance relies on expert demonstrations or fixed intent sets, often failing to generalize to diverse operator habits (inter-operator heterogeneity). Right: Adaptor (Ours) models intent uncertainty via trajectory perturbation and geometric keyframes, conditioning a flow-matching VLA policy to robustly adapt to diverse operator behaviors.

***Abstract*— Assistive teleoperation enhances efficiency via shared control, yet inter-operator variability, stemming from diverse habits and expertise, induces highly heterogeneous trajectory distributions that undermine intent recognition stability. We present Adaptor, a few-shot framework for robust cross-operator intent recognition. The Adaptor bridges the domain gap through two stages: (i) preprocessing, which models intent uncertainty by synthesizing trajectory perturbations via noise injection and performs geometry-aware keyframe extraction; and (ii) policy learning, which encodes the processed trajectories with an Intention Expert and fuses them with the pre-trained vision–language model context to condition an Action Expert for action generation. Experiments on real-world and simulated benchmarks demonstrate that Adaptor achieves state-of-the-art performance, improving success rates and efficiency over baselines. Moreover, the method exhibits low variance across operators with varying expertise, demonstrating robust cross-operator generalization. *The Homepage is available at: https://rainyrobo.github.io/Adaptor/*.**

I. Introduction

Teleoperation systems are foundational to advancing embodied intelligence. By leveraging human multimodal perception and expert decision priors, they enable the collection of demonstration data for training and optimizing robotic generalist policies [1], [2], [3], [4]. In typical systems, inverse kinematics (IK) maps operator-specified end-effector poses to joint-space commands in real time [5], [6]. However, inherent mismatches between human and robot dynamics, particularly under suboptimal operation, can drive joints toward their limits or into singular configurations [7]. Consequently, reducing operator workload while improving teleoperation efficiency and quality remains a central research objective [8], [9], [10].

Assisted teleoperation typically adopts a shared human-robot control paradigm: the user specifies high-level intent, while an autonomous robot takes over low-level control to improve the efficiency and quality of teleoperation [11], [12], [13]. Existing work on user intent inference and execution falls into two categories: (i) methods utilizing predefined intent sets and policy libraries to match user goals with specific behaviors [12], [14], [15], [16], [17], and (ii) data-driven models that map high-dimensional user inputs to low-dimensional, task-specific controls [9], [18], [19], [20], [21], [22], [23]. Despite their effectiveness, both classes of methods rely on trajectory modeling and therefore struggle to generalize intent recognition across operators and skill levels. Even under identical task conditions, differences in user experience and operational routines induce substantial heterogeneity in the distribution of teleoperation trajectories. As a result, existing methods either expand intent-policy libraries or retrain models for each operator, causing data requirements to scale combinatorially with "intent $\times$ operator style$\times$scenario" and thereby significantly increasing the costs of data collection and annotation.

To address these limitations, we propose Adaptor, an assistive teleoperation framework designed to facilitate few-shot adaptation and robust intent recognition across operators with varying experience levels and behavioral profiles. The system operates through a three-stage pipeline. First, Adaptor injects stochastic noise into demonstration trajectories to

†Corresponding author (fanzipei@jlu.edu.cn)
[1]School of Artificial Intelligence, Jilin University. [2]IO-AI TECH.
This work was partially supported by the grants of Jilin Provincial International Cooperation Key Laboratory for Super Smart City and Jilin Provincial Key Laboratory of Intelligent Policing.

form a perturbation distribution, thereby simulating a diverse spectrum of operator behaviors. Second, it employs a keyframe-based intent extraction mechanism to distill critical temporal information, enhancing both the robustness and computational efficiency of intent recognition. Finally, the system encodes these processed trajectories into latent embeddings that condition a Vision-Language-Action (VLA) controller. This controller is trained to recover stable intent representations from perturbations, enabling precise, intent-conditioned policy execution. Our main contributions are summarized as follows:

- We propose Adaptor, a VLA-based shared control framework that balances human intent robustness with VLA efficiency to achieve few-shot cross-operator generalization and reliable task execution.
- We introduce an intent modeling pipeline to tackle cross-operator distribution shift. By synthesizing a perturbation distribution and keyframe extraction, our approach alleviates covariate shift and facilitates the few-shot intent modeling of multi-user intents within an end-to-end policy.
- Across six tasks in simulation and real-world settings on multiple robotic platforms with participants of diverse operational experience, Adaptor achieves state-of-the-art performance, surpassing pure and assisted teleoperation baselines in success rate, completion time, and user satisfaction.

II. Related Work

A. Assisted Teleoperation

Assisted teleoperation offers a pragmatic trade-off between human control and robotic autonomy: it preserves the operator's decision-making flexibility while improving task efficiency [8], [9], [11], [10]. In this paradigm, the operator performs high-level planning and communicates intent, whereas the robot handles low-level closed-loop control and, conditioned on the parsed intent, executes concrete actions [18], [13], [24], [25]. Consequently, the accuracy of intent recognition is a key determinant of overall system performance [26], [27]. Existing approaches fall into two broad categories. The first comprises retrieval-based methods, which collect operators' intents and their associated low-level actions in advance and, at inference time, retrieve the most similar intent-policy pair to assist teleoperation [12], [14], [15], [16], [17]. These methods can further leverage pre-trained Vision-Language models (VLMs) [28] to improve intent inference and align skills more precisely within a library [27]. The second comprises data-driven methods, which jointly learn intent representations and control policies from demonstrations, enabling end-to-end modeling of the intent-policy mapping [9], [18], [19], [20], [21], [22], [23].

However, existing methods face significant challenges in cross-operator generalization. Given the long-tailed distribution of operator behaviors and habits, relying solely on expanding skill libraries or accumulating training data fails to cover these highly personalized and sparse features exhaustively. To address this bottleneck that data scaling alone cannot resolve, we propose a personalized adaptation system based on few-shot learning.

B. Noise Injection to Increase Robustness

Learning assistive teleoperation policies is challenging due to the complex, nonlinear mapping between high-level human intentions and robot dynamics. Standard behavior cloning often suffers from covariate shift, where minor deviations from expert demonstrations induce compounding errors and hazardous states [29], [30]. Drawing on the principle of persistent excitation, researchers can enhance policy robustness and generalization by ensuring that the training data provide broad coverage of the state-action space [31], or by injecting isotropic Gaussian noise into control inputs to induce diverse perturbations [32]. A prevalent strategy involves iteratively executing the robot's current policy while requesting supervisor corrections for the visited states; such online human corrections mitigate policy-induced distribution shift [33], [34], [35], [36]. However, this approach imposes a substantial burden on the supervisor and risks exposing physical hardware to suboptimal or hazardous states. Alternatively, noise injection during data collection can yield diverse demonstrations, with the noise magnitude tuned to approximate the trained policy's error profile (e.g., DART) [37].

Shifting the focus from pure policy correction to user adaptability, we inject stochastic noise into the supervisor policy to simulate a spectrum of teleoperator behaviors. This strategy effectively improves generalization to previously unseen operator behaviors.

C. VLA for Robotics

Vision-Language models (VLMs) pre-trained on internet-scale data have garnered significant attention for their strong generalization and adaptability across diverse applications [38], [39], [40]. Building on these advances, Vision-Language-Action (VLA) systems leverage the semantic priors of pre-trained VLMs to facilitate end-to-end robotic control. This integration enables the development of generalist robot policies capable of adapting to diverse robot embodiments and manipulation tasks [1], [3], [4], [41], [42].

While the autonomous execution capabilities of VLA systems offer the potential to alleviate data collection burdens, current models often lack sufficient generalization and execution stability. To bridge this gap, we propose a VLA-based shared control framework for assistive teleoperation, designed to balance the robustness of human strategies with the efficiency of VLA execution.

III. Problem Formulation

Assistive teleoperation entails both low-level motor control and high-level intent inference. We formalize this process as a Partially Observable Markov Decision Process (POMDP) augmented with a latent intent space, defined by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{P}, \mathcal{Z} \rangle$. Here, $\mathcal{S}$ denotes the unobservable physical state space, $\mathcal{A}$ the action space, and $\mathcal{O}$ the observation space.

At each time step t, the robot receives a composite observation $o_t \in \mathcal{O}$, defined as the tuple $o_t = (s_t^{prop}, o_t^{vis}, L)$. Specifically, s_t^{prop} represents the fully observable proprioceptive state (e.g., gripper status and joint angles), o_t^{vis} denotes the high-dimensional visual observation (RGB images), and L is the natural-language instruction. $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ defines the transition dynamics. The intent $z \in \mathcal{Z}$ is a latent variable modeling the trajectory-level objective, inferred from the human teleoperation trajectory $\xi^h = \{(o_t^h, a_t^h)\}_{t=1}^T$ of horizon T. Consequently, the execution of policy $\pi_\theta(a_t \mid o_t, z)$ is conditioned on the robot's observations and the inferred operator intent.

Formally, our goal is to learn a policy π_θ that approximates an expert policy π_{θ^*}. We analyze the performance gap by decomposing the expected loss into two distinct components:

$$\mathcal{L}_r(\theta) = \underbrace{\mathbb{E}_{p(\xi|\pi_\theta,z)}\mathcal{J}(\theta,\theta^*|\xi) - \mathbb{E}_{p(\xi|\pi_{\theta^*},z)}\mathcal{J}(\theta,\theta^*|\xi)}_{\text{distribution shift}} + \underbrace{\mathbb{E}_{p(\xi|\pi_{\theta^*},z)}\mathcal{J}(\theta,\theta^*|\xi)}_{\text{supervised loss}}, \tag{1}$$

where $p(\xi|\pi, z)$ denotes the trajectory distribution induced by executing policy π conditioned on intent z, and $\mathcal{J}(\theta, \theta^*|\xi)$ measures the discrepancy between π_θ and π_{θ^*} along trajectory ξ. The *supervised loss* represents the standard imitation objective on expert demonstrations. Crucially, the *distribution shift* term quantifies the mismatch between the state distributions visited by the learner versus the expert. In teleoperation, this shift is exacerbated by the high variance of human trajectories ξ^h, stemming from both inter-operator expertise differences and intra-operator stochasticity. Consequently, out-of-distribution (OOD) intents can lead to compounding errors, causing the learned policy to diverge significantly from the expert distribution.

Accordingly, our objective is twofold: (i) to enhance the generalization of intent recognition across varying operator expertise and conditions, thereby mitigating distribution shift; and (ii) to refine policy learning for high-fidelity alignment with expert demonstrations, minimizing the supervised imitation loss.

IV. Method

A. Overview

As illustrated in Fig. 2, the Adaptor framework operates through two distinct phases: *Preprocessing* and *Policy Learning*. During preprocessing, we construct perturbation distributions over teleoperation trajectories and extract keyframes to approximate intent uncertainty, thereby enhancing recognition robustness. In the subsequent training and inference stages, the pipeline orchestrates three core components: a VLM backbone, an Intention Expert, and an Action Expert. Initially, the VLM processes multimodal inputs to extract environmental context. These representations are then utilized by the Intention Expert, which infers latent intentions by fusing trajectory guidance (sourced from preprocessed trajectories during training or coarse human demonstrations during inference) with the scene's semantic context. Finally, the Action Expert integrates these intent predictions with multimodal embeddings to generate comprehensive and precise robot control commands via flow matching.

B. Intention Preprocessing

This subsection addresses the problem of intent *distribution shift* in the error term of Eq. 1 for assisted teleoperation. We introduce two mechanisms: an intent perturbation distribution and a keyframe extraction method to model inter-operator variability in intent and to extract salient intents that suppress local perturbations, thereby mitigating errors induced by the *distribution shift*.

1) Intent Perturbation Distribution: Treating the supervised teleoperation trajectory ξ^* as a direct proxy for the operator's intent ξ integrates seamlessly into standard policy-learning pipelines and obviates additional data collection. However, behavior cloning under this degenerate intent prior exposes the learner only to the expert-induced state distribution, leading to covariate shift at deployment and limited generalization to unseen variations in operator intent.

Inspired by the idea of enhancing policy robustness by injecting noise into expert demonstrations [37], we perturb supervised teleoperation trajectories to construct an intent perturbation distribution that captures inter-operator variability. Concretely, let a supervised demonstration be the teleoperation trajectory $\xi^* = (a_1^*, a_2^*, \ldots, a_T^*)$ collected under an expert policy π_{θ^*}. We model intent variability with a trajectory-level perturbation kernel $q_\psi(\tilde{\xi} \mid \xi^*)$ and adopt a narrow Gaussian tube around ξ^*:

$$\tilde{\xi} \sim q_\psi(\tilde{\xi} \mid \xi^*) \;=\; \mathcal{N}\Big(\tilde{\xi};\, \xi^*,\, K_\psi\Big), \tag{2}$$

where $K_\psi = \mathrm{diag}(\Lambda_1, \ldots, \Lambda_T)$ is taken in block-diagonal form, corresponding to independent per-time perturbations.

From this distribution we draw intent instances with varying perturbation magnitudes and apply behavior cloning via supervised action regression to approximate the expert policy, thereby implicitly recovering the underlying intent from the perturbed trajectories. This procedure expands the state-action coverage during training without relying on a large corpus of suboptimal or failed examples, improving robustness to intent drift and out-of-distribution (OOD) scenarios.

2) Intent Keyframe Extraction: Trajectory segments that are well-approximated by linear interpolation often contain redundant kinematic information regarding the operator's intent. As illustrated in the keyframe extraction module of Fig. 2 for a "pen organization" task, intent is semantically concentrated at subtask boundaries—such as reaching (initiation), grasping (interaction), and releasing (completion)—rather than in the smooth intermediate transition phases. Consequently, preserving every timestep introduces unnecessary computational overhead. Conversely, naive reduction strategies, such as uniform sampling or pooling, lack the geometric sensitivity required to preserve high-frequency control details, potentially missing critical action keyframes such as the exact moment of interaction.

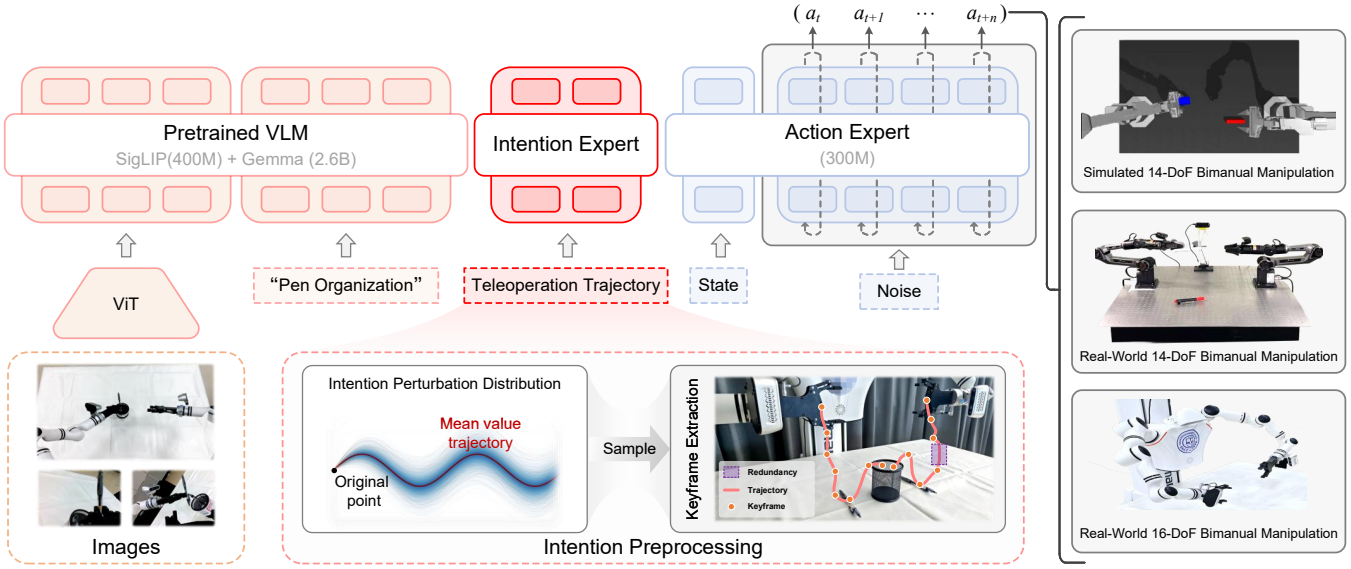

Fig. 2. **Overview of the Adaptor framework.** The architecture comprises two primary phases: (i) Preprocessing, where perturbation distributions and keyframes are extracted to model intent uncertainty; and (ii) Policy Learning. In this phase, the VLM backbone extracts environmental context, while the Intention Expert synthesizes semantic data with preprocessed trajectory guidance to infer latent intent, and the Action Expert employs flow matching to generate precise control commands.

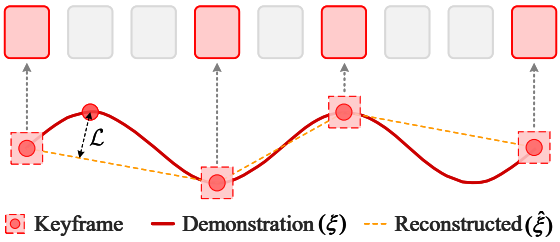

Fig. 3. Schematic of the intent keyframe extraction.

Inspired by [43], we formulate intent extraction as a geometry-aware trajectory compression problem. As illustrated in Fig. 3, given a demonstration trajectory $\xi = \{x_t\}_{t=0}^{T}$, we measure the approximation error of a candidate keyframe subsequence W using the Directed Hausdorff distance. Specifically, for the linear interpolation $\hat{\xi} = f(W)$, the error is defined as:

$$\mathcal{L}(\hat{\xi}, \xi) = \max_{x \in \xi} \min_{\hat{x} \in \hat{\xi}} \ell(x, \hat{x}), \tag{3}$$

where $\min_{\hat{x} \in \hat{\xi}} \ell(x, \hat{x})$ represents the geometric distance from a raw point x to the continuous piecewise-linear trajectory $\hat{\xi}$. We solve for the minimal cardinality set W (satisfying $0 = t_0 < \cdots < t_L = T$) subject to the constraint $\mathcal{L}(\hat{\xi}, \xi) \leq \eta$, thereby ensuring that the worst-case geometric deviation remains within the prescribed tolerance η.

C. VLA-Based Assistive Policy Learning

1) Architecture Implementation: Adaptor adopts a Mixture-of-Transformer architecture integrating three components: a VLM and an Action Expert, both inheriting initial weights from the pre-trained VLA model [1], and a novel decoder-only Intention Expert, implemented in the same manner as the Gemma backbone [28].

The VLM is designed to robustly align the natural language instruction L with visual observations, denoted as $o_t^{vis} = \{I_t^i\}_{i=1}^{N}$. At each timestep t, a SigLIP encoder extracts high-level perceptual features from o_t^{vis}, which are subsequently integrated with language tokens within a shared Transformer space to facilitate deep cross-modal fusion. This architecture explicitly captures the semantic correspondence between task goals and the scene context, yielding consistent representations that serve as a reliable foundation for downstream intent understanding and action execution.

The Intention Expert is designed to model the operator's latent intent z. To enhance robustness against incomplete or interrupted demonstrations encountered during inference, we employ a temporal truncation strategy during training, where expert trajectories are randomly cropped at the end. The preprocessed teleoperation trajectories are first projected into the language embedding space via a linear layer. These trajectory tokens are then concatenated with the VLM features to form a unified context, enabling bidirectional attention for deep semantic alignment and intention recognition.

The Action Expert translates multimodal context into action chunks $A_t = (a_t, a_{t+1}, \ldots, a_{t+H})$. We map the proprioceptive state s_t^{prop} and a flow-matching noise term ϵ into the action embedding space to form queries. These queries drive a cross-modal attention decoder, which attends to the multimodal context (language L, visual observations o_t^{vis}, and intention z) serving as Keys and Values. This architecture ensures that the generated actions are strictly conditioned on both the semantic environment and the specified intention constraints.

2) Training Objective: We adopt a hybrid training strategy to instantiate the policy. To leverage the foundational priors of the pre-trained VLM and Action Expert, we apply Low-Rank Adaptation (LoRA) [44]. Conversely, the Intention Expert undergoes full-parameter training to ensure alignment with the VLM context. Building upon these integrated representations, the policy π_θ employs Conditional Flow Matching (CFM) to model the continuous transformation from a Gaussian prior to the expert action distribution. Specifically, we construct the interpolated action $a_t^\tau = (1-\tau)\epsilon + \tau a_t$ and train π_θ to regress the target vector field $v(a_t^\tau, \tau) = a_t - \epsilon$ via the following objective:

$$\mathcal{L}_{\text{action}} = \mathbb{E}_{t,\tau,\epsilon}\left[\|\pi_\theta(\tau, a_t^\tau, \mathcal{C}) - (a_t - \epsilon)\|^2\right], \qquad (4)$$

where $\mathcal{C} = \{L, o_t^{vis}, s_t^{prop}, z\}$ denotes the conditioning context, comprising language instructions, visual observations, proprioception, and the inferred operator intent z.

V. Experiments

We investigate two research questions: **RQ1**: Does Adaptor enhance manipulation efficiency and quality relative to pure teleoperation, and does it achieve state-of-the-art performance compared to assisted baselines? – Evaluated in Sec. V-B via multi-operator, multi-task comparisons utilizing both objective and subjective metrics. **RQ2**: What are the distinct contributions of each system component to the overall performance? – Quantified in Sec. V-C via ablation studies isolating each component's effect.

A. Experimental Setup

1) System Configurations and Benchmark: We evaluate our method on three robot platforms across six distinct manipulation tasks, as illustrated in Fig. 4.

ALOHA Simulation. Based on the ALOHA setup [45], this simulated environment features two 6-DoF Trossen ViperX arms and a single base camera (14-dimensional action space). We assess two bimanual tasks: *Insertion*, which involves aligning a peg into a block, and *Cube Transfer*, passing a cube from the right arm to the left.

AgileX PIPER. This physical setup comprises two 6-DoF arms monitored by a four-camera system (two wrist, one base, and one low-angle). It operates within a 14-dimensional action space. The tasks include *Pen Uncapping*, coordinating arms to remove a cap, and *Shirt Folding*, folding a T-shirt initially laid flat.

Bimanual Realman. A dual-arm platform equipped with two 7-DoF arms and three cameras (two wrist, one base), utilizing a 16-dimensional action space. We evaluate *Pen Organization*, placing scattered pens into a holder, and *Cube Stacking*, sequentially grasping and stacking cubes vertically.

Teleoperation Interface. As illustrated in Fig. 4 (right), the teleoperation system comprises a Head-Mounted Display (HMD) and handheld controllers. User inputs are mapped to the robot's joint space via Inverse Kinematics (IK). Simultaneously, concurrent multi-view video feeds are streamed to the HMD, facilitating immersive closed-loop control.

2) Baselines: We compare Adaptor against two baselines:

Full Teleop [6]. A direct control method where handheld end-effector poses are mapped to robot joint commands via inverse kinematics, relying solely on manual operator control without autonomous assistance.

HAJL [8]. A shared control framework based on human-agent joint learning, employing a diffusion model to refine human inputs via reverse denoising.

3) Evaluation Metrics: We assess task performance using three metrics:

Teleoperation Quality. The task success rate (%), defined as the proportion of trials successfully completed within a fixed time limit.

Teleoperation Efficiency. The completion time (s), defined as the duration of active human input (excluding autonomous execution time).

User Satisfaction. Subjective feedback assessed using a questionnaire adapted from [27].

4) Participants and Procedures: Eleven healthy participants (3 females, 8 males) were recruited and provided written informed consent. Participants completed a practice session prior to performing the experimental tasks in a randomized order. A total of 30 trials were collected for each task-method combination. To simulate varying levels of operational proficiency, participants were randomly assigned to practice durations of 30, 60, or 120 minutes. Upon completion of each task, participants completed a user satisfaction questionnaire.

B. Comparative Evaluations

1) Comparative Results: Table I presents a quantitative comparison of task success rates and teleoperation times between Adaptor and the baseline methods. Across all six evaluated tasks, Adaptor consistently achieves the highest success rate and the shortest teleoperation time. While model-assisted approaches (both HAJL [8] and Adaptor) generally outperform purely manual teleoperation (Full Teleop) in efficiency and efficacy, Adaptor demonstrates superior robustness. This is notably distinct from HAJL, which lacks explicit modeling of operator intent uncertainty and inter-operator variability, thereby limiting its generalization. Note that the reported teleoperation time strictly measures active human input, excluding the robot's autonomous execution. Unlike synchronous methods (Full Teleop, HAJL), Adaptor decouples demonstration from execution. This asynchronous design holds the potential to facilitate task queuing and Single-Operator-Multi-Robot (SOMR) workflows, allowing operators to efficiently manage parallel objectives.

We attribute these performance gains to three strategic design choices. First, the intent-preprocessing module enhances recognition robustness via stochastic perturbations and keyframe extraction. Second, our random trajectory truncation strategy during training empowers the policy to operate effectively given only partial intent. This capability allows operators to provide concise demonstrations—often executing only a subset of the full trajectory—thereby significantly reducing teleoperation time. Finally, the integration

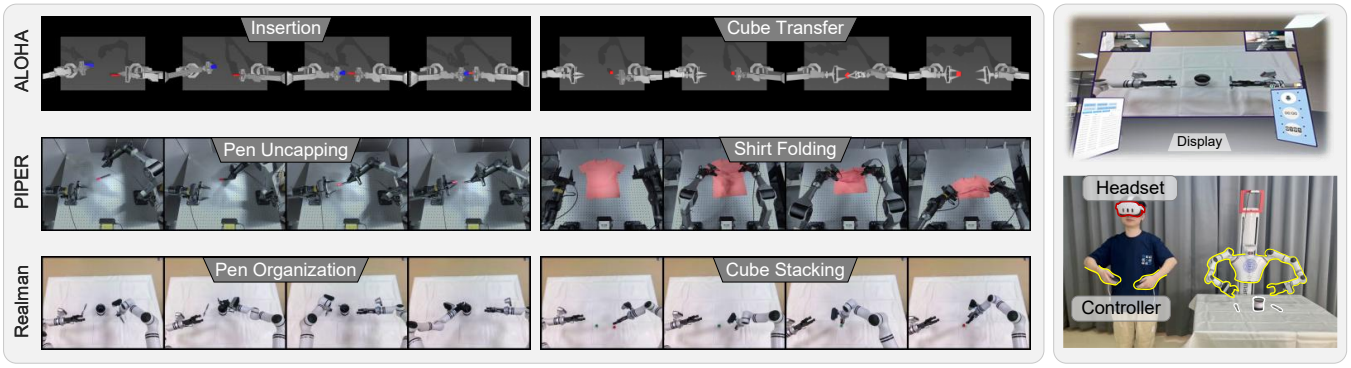

Fig. 4. **Overview of the experimental setup.** Left: Representative tasks across three robotic platforms, including ALOHA simulator (Insertion, Cube Transfer), PIPER (Pen Uncapping, Shirt Folding), and Realman (Pen Organization, Cube Stacking). Right: Schematic of the teleoperation system architecture.

TABLE I
QUANTITATIVE EVALUATION ON REAL-WORLD AND SIMULATED BENCHMARKS.

Robot	**Task**	**Success Rate** (%, ↑)				**Teleoperation Time** (s, ↓)			
		Full Teleop [6]	HAJL [8]	**Adaptor**	Improv.	Full Teleop [6]	HAJL [8]	**Adaptor**	Improv.
ALOHA	Insertion	33.27	41.83	**67.25**	+60.77%	38.77	32.29	**17.92**	-44.50%
	Cube Transfer	63.64	84.64	**91.48**	+8.08%	24.23	19.38	**11.87**	-38.75%
PIPER	Pen Uncapping	45.09	60.75	**89.36**	+47.09%	20.36	15.59	**10.39**	-33.35%
	Shirt Folding	38.00	47.24	**76.10**	+61.09%	23.54	15.73	**11.59**	-26.32%
Realman	Pen Org.	42.91	59.92	**90.27**	+50.65%	31.15	24.49	**17.50**	-28.54%
	Cube Stacking	55.63	69.51	**86.07**	+23.82%	16.13	13.18	**10.34**	-21.55%

Notes: We report the average Success Rate (↑) and Teleoperation Time (↓). Our method (Adaptor) is highlighted in red; "Improv." denotes the relative improvement (%) over the SOTA baseline. The best results are in **bold**.

of a Vision-Language-Action (VLA) backbone improves generalization, ensuring that the inferred intent accurately guides and reinforces end-to-end policy execution.

2) Cross-operator Generalization: Table II presents the evaluation of system robustness across varying levels of operator proficiency (30, 60, and 120 minutes of practice). A positive correlation is observed between practice duration and success rate across all methods. This trend suggests that as operators gain experience, their control inputs increasingly converge towards the expert trajectory distribution, thereby facilitating task completion.

However, Full Teleop and HAJL show performance variations correlated with operator proficiency. In the 30-minute trials, their success rates decrease to 38.43% and 48.76%, respectively, with higher standard deviations (22.08 and 19.20) reflecting increased variability. This decline is attributed to novice trajectories serving as out-of-distribution inputs, which hinders the baselines from mapping actions to the expert policy. In contrast, Adaptor (Ours) maintains a success rate of 83.21% with a standard deviation of 5.69 under the same conditions. These results indicate that the proposed method minimizes the performance disparity between novice and expert users, improving cross-operator generalization.

3) User Satisfaction Results: As illustrated in Fig. 5, Adaptor demonstrates superior performance across nine of the ten metrics, with the notable exception of "Perceived Safety". This distinction stems from the fundamental difference in interaction paradigms. Adaptor operates via a one-shot demonstration followed by an autonomous execution workflow. While this paradigm substantially reduces operator workload, the absence of real-time intervention capabilities during the autonomous phase diminishes the operator's sense of agency. Consequently, this lowers perceived safety, reflecting a trade-off often observed in high-autonomy systems [46].

In contrast, HAJL employs an iterative correction mechanism. However, its effectiveness is compromised by unstable intent inference and the need for frequent interruptions. These factors introduce temporal inefficiencies and degrade user experience, resulting in significantly lower scores for "Trust" and "Willingness to Reuse". Finally, while Full Teleop imposes the highest physical load due to the requirement for continuous manual input, its inherent real-time controllability fosters a higher sense of safety compared to

TABLE II
CROSS-OPERATOR GENERALIZATION ANALYSIS.

Method	**Success Rate** (%) ↑			**SD** ↓
	30 min	**60 min**	**120 min**	
Full Teleop [6]	38.43	62.17	82.54	22.08
HAJL [8]	48.76	71.32	86.95	19.20
Adaptor (Ours)	**83.21**	**89.85**	**94.53**	**5.69**

Notes: Success Rate (%) is averaged across three proficiency levels for each practice duration. The SD serves to quantify generalization, where a lower value indicates superior consistency across these levels.

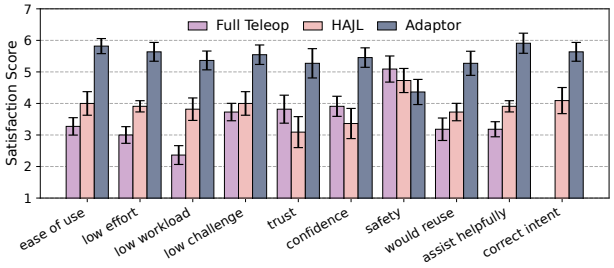

Fig. 5. **Quantitative Analysis of User Satisfaction.** Mean satisfaction scores derived from questionnaires administered after participants completed 30 trials for each task–method combination. Data are averaged across all tasks, with error bars representing the standard deviation (SD).

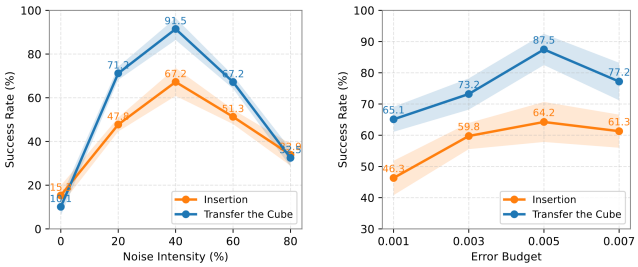

Fig. 6. **Ablation study results**. Analysis of the proposed intent perturbation distribution and keyframe extraction on the Insertion and Cube Transfer tasks in the ALOHA simulation environment.

semi-autonomous approaches.

C. Ablation Studies

1) Ablation Study on Noise Injection: To evaluate whether injecting noise into the intent representation improves trajectory diversity and generalization across operators, we conduct an ablation study on two tasks in the ALOHA simulation environment, as shown in Fig. 6 (left). The x-axis denotes the noise level (0 = no noise, i.e., the demonstrator trajectory is used directly as the intent), and the y-axis reports the task success rate. The results exhibit a clear inverted-U trend. With near-zero noise, the model overfits to the demonstrator's style and fails to generalize, yielding the lowest success rates. Conversely, excessive noise obscures the underlying intent and hinders recovery, also degrading performance. At a moderate noise level, the perturbation acts as an effective regularizer, encouraging task-relevant, operator-invariant intent representations and achieving the highest success.

2) Ablation Study on Keyframe Extraction: Fig. 6 (right) presents the success rates of two tasks in the ALOHA simulation environment as a function of trajectory reconstruction error. The horizontal axis denotes the reconstruction error (smaller values indicate a closer match between the reconstructed and original trajectories, i.e., denser keyframes), and the vertical axis denotes the task success rate. Both tasks exhibit a characteristic inverted-U relationship: when the Error Budget (η) is too large, the extracted keyframes are overly sparse, yielding an underspecified trajectory representation that fails to capture the user's latent intent; conversely, when η is too small, excessively dense keyframes introduce redundant waypoints and accumulate noise, increasing planning and control overhead and inducing overfitting, which likewise reduces success rates. Overall, a moderate η strikes a balance between representational fidelity and generalization, producing the highest success rates on both tasks.

VI. Conclusion

We present Adaptor, an assistive teleoperation framework designed to achieve intent generalization across diverse operators. By integrating intent uncertainty simulation during the preprocessing phase with explicit intent modeling in end-to-end training, Adaptor effectively addresses the challenge of recognizing out-of-distribution (OOD) intents stemming from long-tail operator behaviors. Extensive evaluations across six manipulation tasks demonstrate that Adaptor significantly outperforms state-of-the-art (SOTA) methods, achieving a 41.92% increase in average success rate and a 32.17% reduction in teleoperation time. Furthermore, a significant reduction in the standard deviation of success rates across operators of varying proficiency validates the system's superior cross-operator generalization capabilities. User studies further confirm that our approach yields substantially higher user satisfaction.

While Adaptor outperforms assisted teleoperation baselines across varying experience levels, intent recognition remains susceptible to errors when operator behavior diverges significantly from the demonstrator's distribution (e.g., with untrained novices). Furthermore, although our assistance operates autonomously to markedly reduce operator workload, post-study feedback indicates that concerns regarding system safety persist. In future work, we propose adopting a hybrid supervised-reinforcement fine-tuning strategy designed to facilitate controlled exploration of out-of-distribution (OOD) intents. This will allow us to strictly bolster safety and trust while maintaining minimal operator intervention.

References

[1] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[2] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu, "Rdt-1b: a diffusion foundation model for bimanual manipulation," *arXiv preprint arXiv:2410.07864*, 2024.

[3] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[4] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*. PMLR, 2023, pp. 2165–2183.

[5] A. Iyer, Z. Peng, Y. Dai, I. Guzey, S. Haldar, S. Chintala, and L. Pinto, "Open teach: A versatile teleoperation system for robotic manipulation," *arXiv preprint arXiv:2403.07870*, 2024.

[6] K. Zakka, "Mink: Python inverse kinematics based on MuJoCo," Feb. 2026. [Online]. Available: https://github.com/kevinzakka/mink

[7] A. Phung, G. Billings, A. F. Daniele, M. R. Walter, and R. Camilli, "A shared autonomy system for precise and efficient remote underwater

manipulation," *IEEE Transactions on Robotics*, vol. 40, pp. 4147–4159, Jan. 2024.

[8] S. Luo, Q. Peng, J. Lv, K. Hong, K. R. Driggs-Campbell, C. Lu, and Y.-L. Li, "Human-agent joint learning for efficient robot manipulation skill acquisition," *arXiv preprint arXiv:2407.00299*, 2024.

[9] T. Yoneda, L. Sun, G. Yang, B. Stadie, and M. Walter, "To the noise and back: Diffusion for shared autonomy," *arXiv preprint arXiv:2302.12244*, 2023.

[10] S. Liu, A. Hasan, K. Hong, R. Wang, P. Chang, Z. Mizrachi, J. Lin, D. L. McPherson, W. A. Rogers, and K. Driggs-Campbell, "Dragon: A dialogue-based robot for assistive navigation with visual language grounding," *IEEE Robotics and Automation Letters*, vol. 9, no. 4, pp. 3712–3719, 2024.

[11] A. Padmanabha, J. Gupta, C. Chen, J. Yang, V. Nguyen, D. J. Weber, C. Majidi, and Z. Erickson, "Independence in the home: A wearable interface for a person with quadriplegia to teleoperate a mobile manipulator," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 542–551.

[12] C. Brooks and D. Szafir, "Balanced information gathering and goal-oriented actions in shared autonomy," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2019, pp. 85–94.

[13] M. Selvaggio, M. Cognetti, S. Nikolaidis, S. Ivaldi, and B. Siciliano, "Autonomy in physical human-robot interaction: A brief survey," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7989–7996, 2021.

[14] S. Jain and B. Argall, "Probabilistic human intent recognition for shared autonomy in assistive robotics," *ACM Transactions on Human-Robot Interaction*, vol. 9, no. 1, 2020.

[15] S. Javdani, H. Admoni, S. Pellegrinelli, S. S. Srinivasa, and J. A. Bagnell, "Shared autonomy via hindsight optimization for teleoperation and teaming," *The International Journal of Robotics Research*, vol. 37, no. 7, pp. 717–742, 2018.

[16] A. Jonnavittula and D. P. Losey, "I know what you meant: Learning human objectives by (under) estimating their choice set," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 2747–2753.

[17] B. A. Newman, R. M. Aronson, S. S. Srinivasa, K. Kitani, and H. Admoni, "HARMONIC: A multimodal dataset of assistive human–robot collaboration," *The International Journal of Robotics Research*, vol. 41, no. 1, pp. 3–11, 2022.

[18] S. Chen, J. Gao, S. Reddy, G. Berseth, A. D. Dragan, and S. Levine, "Asha: Assistive teleoperation via human-in-the-loop reinforcement learning," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 7505–7512.

[19] M. Zhao, R. Simmons, H. Admoni, and A. Bajcsy, "Conformalized teleoperation: Confidently mapping human inputs to high-dimensional robot actions," *arXiv preprint arXiv:2406.07767*, 2024.

[20] Y. Cui, S. Karamcheti, R. Palleti, N. Shivakumar, P. Liang, and D. Sadigh, "No, to the right: Online language corrections for robotic manipulation via shared autonomy," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2023, pp. 93–101.

[21] A. Jonnavittula and D. P. Losey, "Learning to share autonomy across repeated interaction," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 1851–1858.

[22] M. Zurek, A. Bobu, D. S. Brown, and A. D. Dragan, "Situational confidence assistance for lifelong shared autonomy," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 2783–2789.

[23] A. Jonnavittula, S. A. Mehta, and D. P. Losey, "Sari: Shared autonomy across repeated interaction," *ACM Transactions on Human-Robot Interaction*, vol. 13, no. 2, pp. 1–36, 2024.

[24] J. Gu, S. Kirmani, P. Wohlhart, Y. Lu, M. G. Arenas, K. Rao, W. Yu, C. Fu, K. Gopalakrishnan, Z. Xu *et al.*, "Rt-trajectory: Robotic task generalization via hindsight trajectory sketches," *arXiv preprint arXiv:2311.01977*, 2023.

[25] P. Sundaresan, Q. Vuong, J. Gu, P. Xu, T. Xiao, S. Kirmani, T. Yu, M. Stark, A. Jain, K. Hausman *et al.*, "Rt-sketch: Goal-conditioned imitation learning from hand-drawn sketches," in *8th Annual Conference on Robot Learning*, 2024.

[26] G. Hoffman, T. Bhattacharjee, and S. Nikolaidis, "Inferring human intent and predicting human action in human–robot collaboration," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 7, no. 1, pp. 73–95, 2024.

[27] H. Liu, R. Shah, S. Liu, J. Pittenger, M. Seo, Y. Cui, Y. Bisk, R. Martín-Martín, and Y. Zhu, "Casper: Inferring diverse intents for assistive teleoperation with vision language models," *arXiv preprint arXiv:2506.14727*, 2025.

[28] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière *et al.*, "Gemma 3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2503.19786

[29] S. A. Mehta, Y. U. Ciftci, B. Ramachandran, S. Bansal, and D. P. Losey, "Stable-bc: Controlling covariate shift with stable behavior cloning," *IEEE Robotics and Automation Letters*, 2025.

[30] E. Baek, K. Park, J. Kim, and H.-S. Kim, "Unexplored faces of robustness and out-of-distribution: Covariate shifts in environment and sensor domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 294–22 303.

[31] W. Liu, G. Duan, M. Hou, and H. Kong, "Robust adaptive control of high-order fully-actuated systems: Command filtered backstepping with concurrent learning," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 71, no. 12, pp. 5780–5791, 2024.

[32] M. Green and J. B. Moore, "Persistence of excitation in linear systems," *Systems & control letters*, vol. 7, no. 5, pp. 351–360, 1986.

[33] S. Ross and D. Bagnell, "Efficient reductions for imitation learning," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 661–668.

[34] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 627–635.

[35] Y. Wang, L. Wang, Y. Du, B. Sundaralingam, X. Yang, Y.-W. Chao, C. Pérez-D'Arpino, D. Fox, and J. Shah, "Inference-time policy steering through human interactions," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 15 626–15 633.

[36] M. Kelly, C. Sidrane, K. Driggs-Campbell, and M. J. Kochenderfer, "Hg-dagger: Interactive imitation learning with human experts," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8077–8083.

[37] M. Laskey, J. Lee, R. Fox, A. Dragan, and K. Goldberg, "Dart: Noise injection for robust imitation learning," in *Conference on robot learning*. PMLR, 2017, pp. 143–156.

[38] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[39] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[40] P. Gao, J. Han, R. Zhang, Z. Lin, S. Geng, A. Zhou, W. Zhang, P. Lu, C. He, X. Yue *et al.*, "Llama-adapter v2: Parameter-efficient visual instruction model," *arXiv preprint arXiv:2304.15010*, 2023.

[41] J. Liu, H. Chen, P. An, Z. Liu, R. Zhang, C. Gu, X. Li, Z. Guo, S. Chen, M. Liu *et al.*, "Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model," *arXiv preprint arXiv:2503.10631*, 2025.

[42] Physical Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai *et al.*, "$\pi_{0.5}$: A vision-language-action model with open-world generalization," *arXiv preprint arXiv:2504.16054*, 2025.

[43] K. Darvish, L. Penco, J. Ramos, R. Cisneros, J. Pratt, E. Yoshida, S. Ivaldi, and D. Pucci, "Teleoperation of humanoid robots: A survey," *IEEE Transactions on Robotics*, vol. 39, no. 3, pp. 1706–1727, 2023.

[44] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *Iclr*, vol. 1, no. 2, p. 3, 2022.

[45] T. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *RSS*, vol. abs/2304.13705, 2023.

[46] M. A. Collier, R. Narayan, and H. Admoni, "The sense of agency in assistive robotics using shared autonomy," in *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2025, pp. 880–888.