IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

CaFe-TeleVision: A Coarse-to-Fine Teleoperation System with Immersive Situated Visualization for Enhanced Ergonomics

Zixin Tang, Yiming Chen, Quentin Rouxel, Dianxi Li, Shuang Wu, and Fei Chen, *Senior Member, IEEE*

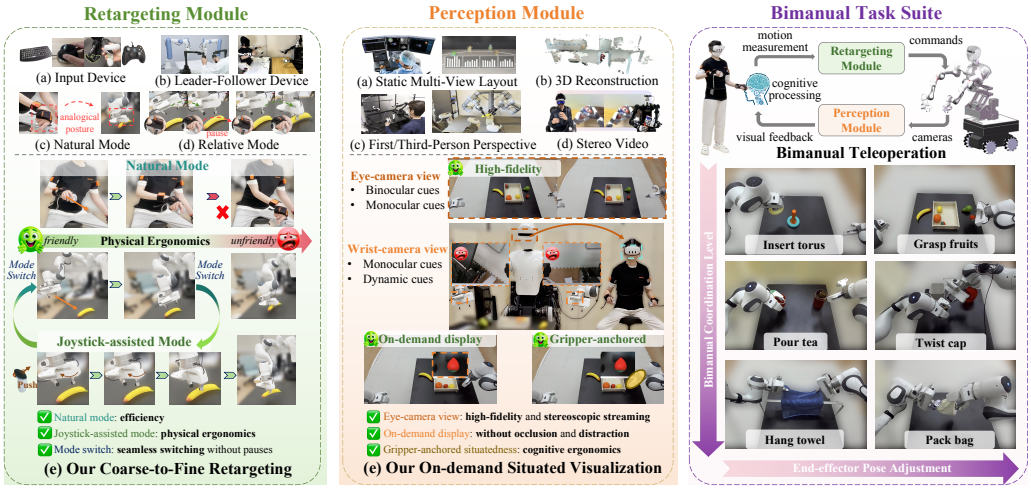


Fig. 1: *Left and Middle*: Comparison of various teleoperation system designs for retargeting and perception module, respectively. Unlike conventional approaches *(a)-(d)*, **CaFe-TeleVision** focuses on jointly enhancing efficiency, physical ergonomics, and cognitive ergonomics. *Left (e)* presents our coarse-to-fine retargeting mechanism that preserves physical ergonomics via seamless mode switching. *Middle (e)* shows our novel on-demand situated visualization technique to reduce eye focus shift, visual distraction, and occlusion, thereby improving cognitive ergonomics during multi-view processing. *Right*: Six tasks for evaluating teleoperation systems, spanning different bimanual coordination levels and end-effector pose adjustment requirements.

***Abstract*— Teleoperation presents a promising paradigm for remote control and robot proprioceptive data collection. Despite recent progress, current teleoperation systems still suffer from limitations in efficiency and ergonomics, particularly in challenging scenarios. In this paper, we propose CaFe-TeleVision, a coarse-to-fine teleoperation system with immersive situated visualization for enhanced ergonomics. At its core, a coarse-to-fine control mechanism is proposed in the retargeting module to bridge workspace disparities, jointly optimizing efficiency and physical ergonomics. To stream immersive feedback with adequate visual cues for human vision systems, an on-demand situated visualization technique is integrated in the perception module, which reduces the cognitive load for multi-view processing. The system is built on a humanoid collaborative robot and validated with six challenging bimanual manipulation tasks. User study among 24 participants confirms that CaFe-TeleVision enhances ergonomics with statistical significance, indicating a lower task load and a higher user acceptance during teleoperation. Quantitative results also validate the superior performance of our system across six tasks, surpassing comparative methods by up to 28.89% in success rate and accelerating by 26.81% in completion time. Project webpage: https://clover-cuhk.github.io/cafe_television/**

Manuscript received 21 July 2025; revised 17 October 2025; accepted 14 November 2025. This paper was recommended for publication by Editor Ki-Uk Kyung upon evaluation of the Associate Editor and Reviewers' comments. This work was supported in part by the Research Grants Council of the Hong Kong SAR under Grant 14211723, 14222722, 24209021 and C7100-22GF, in part by CUHK & HUAWEI Foundation Models and Interactive Intelligence Innovation Laboratory TH2520452 and in part by InnoHK of the Government of Hong Kong via the Hong Kong Centre for Logistics Robotics. *(Corresponding authors: Fei Chen)*

Zixin Tang, Yiming Chen, Quentin Rouxel, Dianxi Li, and Fei Chen are with the Department of Mechanical and Automation Engineering, T-Stone Robotics Institute, The Chinese University of Hong Kong, Hong Kong SAR (email: zxtang@mae.cuhk.edu.hk, ymchen@mae.cuhk.edu.hk, quentinrouxel@cuhk.edu.hk, dxli@mae.cuhk.edu.hk, f.chen@ieee.org).

Shuang Wu is with Huawei Hong Kong Research Center (email: wushuangust@gmail.com).

Digital Object Identifier (DOI): see top of this page.

I. Introduction

Teleoperation, one of the earliest research areas in robotics, offers a promising paradigm to allow human operators to remotely control robots to perform various tasks in different environments, such as medical surgery and space exploration [1]. With the rise of data-driven skill learning [2], [3], [4], [5], teleoperation systems have gained prominence as an effective demonstration collection approach [6], [7], [8], [9].

Typically, teleoperation systems include two modules: a retargeting module that maps human motions to robot control references, and a perception module that obtains perceptual feedback from robot side [10]. In retargeting module, task-space mapping [8], [11] converts human hand poses into target poses for robot end-effectors, allowing more flexibility than joint-space mapping [7]. In perception module, many studies rely on 2D monitors [12] or first/third-person perspectives [13], [14] for feedback, which either lack stereoscopic perception [15], [16] or hinder remote teleoperation. To mitigate these limitations, streaming stereo video to a Virtual Reality (VR) head-mounted display (HMD) provides an effective solution [17], [18], [19].

Despite recent progress, two challenges remain inadequately addressed in existing approaches, which compromise teleoperation

performance in efficiency and ergonomics. First, morphological and kinematic discrepancies between humans and robots introduce the workspace mismatch problem in task-space mapping. This issue is particularly prominent in orientation workspace, due to the limited rotational range of the operator's anatomical joints (e.g., ulnar/radial deviation) [20]. Improper retargeting induces muscle strain in teleoperation, adverse to physical ergonomics [21]. Second, stereo video streams from robot eye cameras fail to provide adequate visual cues for human vision systems, such as dynamic cues (motion parallax) [22]. This limitation is troubled by occlusion-induced blind spots, leading to poor cognitive ergonomics [21] and ultimately impairing vision-based decision-making.

We aim to bridge the challenges in a practical and easily deployable manner. To this end, we propose CaFe-TeleVision, a coarse-to-fine teleoperation system with immersive situated visualization for enhanced ergonomics, as shown in Fig 1. At its core, a coarse-to-fine control mechanism is proposed to bridge workspace disparities through two complementary modes: natural mode and joystick-assisted mode. Specifically, the natural mode follows the pattern of scaling position and aligning orientation to retarget human wrist motions into analogous target poses for end-effectors. This intuitive analogy allows operators to perceive robots as kinematic avatars for efficient operation. When encountering kinematic limits due to orientation mismatch, they can seamlessly switch to the joystick-assisted mode for incremental refinement. The dual modes jointly improve efficiency and physical ergonomics.

In addition, stereoscopic visual feedback from the eye camera is streamed in real time as the primary display, complemented by two wrist cameras. Unlike static multi-view layout, an immersive situated visualization technique is integrated in perception module. Specifically, situated visualization is an emerging concept within visualization, which means displaying data in spatial proximity with its contextual referent (i.e., situatedness) [23]. To reduce eye focus shift, visual distraction, and occlusion, our system implements on-demand display and gripper-anchored situatedness. The operator can control the visibility of wrist views through signals of VR controllers and visualize them spatially coupled to the gripper. These designs enhance cognitive ergonomics during multi-view processing.

The system is evaluated on a humanoid collaborative robot with six bimanual tasks, covering various collaborative levels and operational difficulty (see Fig. 1). Experiments in the user study confirm that CaFe-TeleVision significantly improves ergonomics, with a lower task load and a higher user acceptance during teleoperation. Quantitative results also validate its superior performance, boosting success rates by up to 28.89% and efficiency by 26.81%.

II. Related Works

A. Measurement and Feedback in Teleoperation

Teleoperation systems utilize various approaches to measure human motion information and transmit perceptual feedback [1]. Table I summarizes their characteristics, focusing mainly on kinematic measurement and visual feedback. Fig. 1 presents examples. A more comprehensive review can be found in [10]. Typically, motion measurement methods use user interaction or measurement algorithms for estimation, which can be classified into five categories: input device (e.g., keyboard, mouse), Leader-Follower device (e.g., exoskeleton, Da Vinci), vision-based detector (e.g., OpenPose [24]), optical tracking (e.g., OptiTrack), and IMU sensor (e.g., Xsens [25]). In comparison, IMU sensors show relatively higher performance with respect to intuitiveness, portability, occlusion robustness, and precision. For visual feedback, operators always observe the workspace through GUI monitors, first/third-person perspective, or VR HMDs (e.g., Meta Quest). Recently, VR HMDs are sprouting as an effective solution, replicating immersive stereo vision via 3D reconstruction [26], [27] or stereo streaming [6], [19], where the latter maintains higher fidelity.

TABLE I: Characteristics of motion measurement and visual feedback approaches.

Motion Measurement	Intuitive	Portable	Occlusion-robust	High-precision
Input Device	✗	✓	✓	✗
Leader-Follower Device	✓	✗	✓	✓
Vision-based Detector	✓	✓	✗	✗
Optical Tracking	✓	✗	✗	✓
IMU Sensor	✓	✓	✓	✓
Visual Feedback	**Immersive**	**Stereoscopic**	**Spatial-unrestricted**	**High-fidelity**
GUI Monitor	✗	✗	✓	✓
Third-person perspective	✗	✓	✗	✓
First-person perspective	✓	✓	✗	✓
3D Reconstruction	✓	✓	✓	✗
Stereo Stream	✓	✓	✓	✓

B. Retargeting Module in Teleoperation

The retargeting module maps measured human motions to robot control commands. Joint-space mapping [7], [6] transmits the joint configurations of the leader arms to the follower manipulators, which requires similar morphologies between leaders and followers. Task-space mapping [28], [29], [8], [11], [30], [31], [18], [13], [32] widely measures human hand poses as motion references for robot end-effectors, supporting cross-morphological teleoperation. Later, advanced control strategies can be integrated to solve low-level commands [33]. Depending on the selection of the operator's motion reference frames, it can be further classified into natural mode and relative mode. The natural mode [28], [8] defines fixed base frames (e.g., shoulders) and computes the hand poses based on these frames. In contrast, the relative mode incorporates a trigger to enable/disable retargeting, setting the base frames to the last-step hand poses. Control commands are updated only when enabling retargeting. Previous work with natural mode [8], [19] tuned a fixed scaling factor for position mapping while leaving more challenging orientation mapping as identical alignment. The posture analogy is intuitive for operators to view the robot as their kinematic avatar, which facilitates high efficiency. However, improper factors and overlooking of orientation mismatch sacrifice physical ergonomics. Some systems use relative mode to relieve the problem at a cost of efficiency, which allows operators to adjust hand postures during pauses between disabling and enabling [13]. In CaFe-TeleVision, we propose a coarse-to-fine retargeting mechanism that jointly optimizes efficiency and physical ergonomics.

C. Perception Module in Teleoperation

Prior research has shown that human vision systems rely on a combination of visual cues to perceive (e.g., depth) and then cognitive processing, including proprioceptive cues, monocular

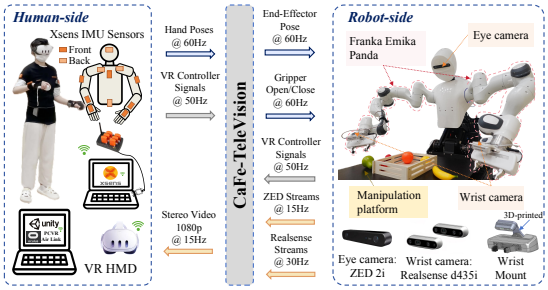


Fig. 2: Workflow of **CaFe-TeleVision** teleoperation system. The retargeting flow (⇒) supports real-time control, retargeting human motions measured by Xsens and VR controllers to robot commands. The perception flow (⇐) streams multi-view stereo video from robot side to a VR head-mounted display (HMD) through a GPU-accelerated Unity application, ensuring high-resolution, real-time visual feedback.

cues, binocular cues, and dynamic cues [22]. To enhance immersive feelings, Cheng et al. [19] streamed stereo video from a robot's eye camera to an HMD to reproduce binocular visual cues. However, limited dynamic cues (motion parallaxes) also hamper cognitive ergonomics, restricted by occlusion-induced blind spots. NimbRo system [6] mitigated it by using a 6-DoF neck arm, which allows large-range rotational and translational changes of observation viewpoints. Despite the performance, this superiority comes at the expense of increased system complexity and higher costs. Recent approaches based on 3D reconstruction [26], [27] offer controllable scenes that support immersive and active operator observation. However, these methods face a fundamental trade-off between high fidelity and computational efficiency. Alternatively, multi-view visualization complements visual cues via increased viewpoints, such as close-up observations. Unlike static multi-view layout, situated visualization as an emerging technique highlights displaying data spatially coupled with its context, which is conducive to cognitive processing [34]. In CaFe-TeleVision, we use two wrist cameras and integrate situated visualization [23], [35] to enhance cognitive ergonomics.

III. Teleoperation System

A. Overview

CaFe-TeleVision is a practical and easily deployable VR-based teleoperation system, built on a humanoid collaborative robot. As shown in Fig. 2, the workflow forms a retargeting-perception loop between human and robot sides. Fig. 3 presents core designs in our system, including a coarse-to-fine retargeting module and an immersive perception module, along with VR controller interface utilized in them to enhance ergonomics. The complete system implementation is available online on the project's website.

In retargeting flow, we use 11 Xsens sensors [25] to capture high-accuracy hand poses in SE(3) relative to shoulder frame, which directly reflects the target transformation references in human arm workspace. The poses are transmitted via UDP packets at 60 Hz. Simultaneously, VR controllers signals are input at 50 Hz. After retargeting, the output 7-DoF commands per arm are published at 60 Hz via ROS, then converted into 7-DoF joint torques through a model-predictive impedance controller [36] for Franka arm, with gripper open/close command. The controller

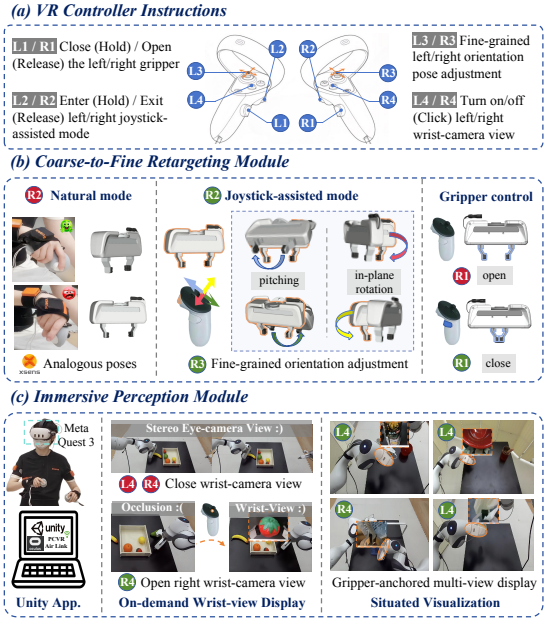


Fig. 3: Framework of **CaFe-TeleVision** teleoperation system. At its core, (a) depicts the functionality of VR controller interface, utilized for assisting teleoperation. (b) shows our coarse-to-fine retargeting module with natural mode for efficient motions and joystick-assisted mode for enhanced physical ergonomics via orientation refinements. (c) illustrates our immersive perception module with on-demand situated visualization technique to improve cognitive ergonomics during multi-processing. Button states: inactivate (●) and active (●).

operates at 500 Hz, integrated with constraints on position, speed, control, and contact torques, which guarantees stable, smooth, and real-time control.

In perception flow, a ZED 2i stereo camera is mounted as the robot's eye, providing a global view of the manipulation platform. Two RealSense D435i cameras are mounted on the flanges with 3D-printed adapters. These video streams are processed in Unity application running on a GPU-accelerated laptop, supporting 1080p@15 Hz real-time visual feedback.

B. Coarse-to-Fine Retargeting Module

The retargeting module maps operator hand poses to robot commands. However, human-robot morphological and kinematic discrepancies inherently cause the workspace mismatch. For example, the human wrist has a limited range of motion, making certain rotations difficult or physically uncomfortable for the operator. In contrast, the robot with a different structural design may handle those motions more easily. To this end, we propose a coarse-to-fine mechanism with two modes: the natural mode for efficient but coarse-grained motions and the joystick-assisted mode for ergonomic and fine-grained adjustments. The dual modes complement each other and jointly optimize efficiency and physical ergonomics.

1) Natural mode: The mode follows the pattern of scaling position and aligning orientation to map human wrist motions into analogous target 6-DoF poses for end-effectors. Specifically, the workspaces of both hands are modeled as separate spheres, where the origin is on the operator's shoulder and the radius is

arm length, denoted as r_H. The intersection area depends on the origin distance of spheres, defined as d_H. To achieve full coverage and maintain physical ergonomics in bimanual manipulation, we calculate the scaling factor $\boldsymbol{s}$ for each axis based on the mismatch in radii and origin distances:

$$\boldsymbol{s}=\begin{bmatrix}\frac{r_C}{r_H} & \max(\frac{r_C}{r_H},\frac{d_C}{d_H}) & \frac{r_C}{r_H}\end{bmatrix}^T, \tag{1}$$

where r_C is the workspace radius of Franka, and d_C is its origin distance. Unlike anthropomorphic arms, the origin of Franka lies in the second joint. We thereby transform the retargeted position from the origin to the tilted arm base via ${}^i\boldsymbol{T}_w^{ab}$, where $i\in\{\text{L,R}\}$ denotes the arm side. The retargeted position ${}^i\boldsymbol{p}_{ee}^{ab}(t)$ for the end-effector is computed as:

$${}^i\tilde{\boldsymbol{p}}_{ee}^{ab}(t)={}^i\boldsymbol{T}_w^{abi}\tilde{\boldsymbol{p}}_{ee}^{w}(t),\quad {}^i\boldsymbol{p}_{ee}^{w}(t)=\boldsymbol{s}\circ{}^i\boldsymbol{p}_{hand}^{w}(t), \tag{2}$$

where human hand position ${}^i\boldsymbol{p}_{hand}^{w}(t)$ based on the shoulder frame is obtained from Xsens in real time, $\circ$ is Hadamard product, and $\tilde{\boldsymbol{p}}$ refers to homogeneous coordinate format.

We align the 3-DoF orientation via ${}^i\boldsymbol{R}_{ee}^{hand}$, identically mapping hand to the approaching direction, the rotation in ulnar/radial deviation to in-plane rotation of the grippers and extension/flexion to pitching. Then, the retargeted orientation ${}^i\boldsymbol{R}_{ee}^{ab}(t)$ related to the arm base frame is defined as:

$${}^i\boldsymbol{R}_{ee}^{ab}(t)={}^i\boldsymbol{R}_{w}^{abi}\boldsymbol{R}_{hand}^{w}(t){}^i\boldsymbol{R}_{ee}^{hand}, \tag{3}$$

where ${}^i\boldsymbol{R}_{hand}^{w}(t)$ is the orientation of hand in real time. Benefiting from intuitiveness, this mode supports coarse-grained but high-amplitude motions, ensuring high efficiency.

2) Joystick-assisted mode: When facing non-ergonomic risk in natural mode, operators can seamlessly switch to joystick-assisted mode by holding L2/R2, as exemplified in Fig. 3 (b). Fine-grained pose adjustments can be applied using 2-DoF thumbsticks, one for in-plane rotation and the other for pitching. And the hypothesis for fixing the approaching direction is that humans have greater freedom to adjust their wrists to feasible directions in natural mode. Note that position mapping still follows the natural mode.

When holding the R2 button, the orientation of right gripper stops aligning with Eq. 3. Instead, it maintains a continuous 2-DoF rotation following the scrolling of R3. Formally, let ${}^iu_1,{}^iu_2\in[-1,1]$ denote the normalized 2-DoF signal from the i-side thumbstick, where ${}^iu_1$ denotes horizontal deflection and ${}^iu_2$ is vertical deflection. To enhance action-perception consistency, we implement a transparent spatial mapping: horizontal inputs (left-right) control gripper's in-plane rotation and vertical inputs (up-down) adjust its pitching. Then, ${}^iu_1,{}^iu_2$ are retargeted to incremental angular change, defined as:

$${}^i\theta_1=s_1{}^iu_1,\quad {}^i\theta_2=s_2{}^iu_2, \tag{4}$$

where ${}^i\theta_1$, ${}^i\theta_2$ denote in-plane angle and pitch angle, respectively. s_1 and s_2 are rotational scaling factors. Assuming i-side joystick-assisted mode engaged at t_0, ${}^i\boldsymbol{R}_{ee}^{ab}(t_0)$ is recorded as the initial orientation, then updated as:

$${}^i\boldsymbol{R}_{ee}^{ab}(t)\leftarrow{}^i\boldsymbol{R}_{ee}^{ab}(t)\Delta\boldsymbol{R}({}^i\theta_2(t))\Delta\boldsymbol{R}({}^i\theta_1(t)), \tag{5}$$

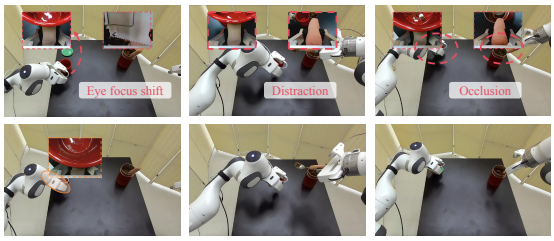


Fig. 4: Comparison of multi-view display between "always on" static layout (top) and on-demand situated visualization (bottom). The bottom pattern (ours) leads to low cognitive load, in terms of eye focus shift, distraction, and occlusion.

where $\Delta\boldsymbol{R}({}^i\theta_1(t))$ and $\Delta\boldsymbol{R}({}^i\theta_2(t))$ (see Eq. 6) are rotation for in-plane direction and pitching, respectively.

$$\begin{aligned}\Delta\boldsymbol{R}({}^i\theta_1(t))&=\begin{bmatrix}\cos({}^i\theta_1(t)) & -\sin({}^i\theta_1(t)) & 0\\ \sin({}^i\theta_1(t)) & \cos({}^i\theta_1(t)) & 0\\ 0 & 0 & 1\end{bmatrix},\\ \Delta\boldsymbol{R}({}^i\theta_2(t))&=\begin{bmatrix}\cos({}^i\theta_2(t)) & 0 & \sin({}^i\theta_2(t))\\ 0 & 1 & 0\\ -\sin({}^i\theta_2(t)) & 0 & \cos({}^i\theta_2(t))\end{bmatrix}\end{aligned} \tag{6}$$

The integration of the joystick-assisted mode in retargeting module allows operators to maintain ergonomic postures while gradually adjusting to feasible poses. When L2/R2 is released, ${}^i\boldsymbol{R}_{ee}^{ab}(t)$ is interpolated to real-time human hand poses via SLERP algorithm, smoothly rotating gripper towards targets at a constant angular velocity. The interpolation automatically terminates if the rotation distance between gripper's current orientation and the hand orientation is below the threshold. Then, the identical alignment in natural mode is recovered immediately.

C. Immersive Perception Module

Recent work leverages VR HMDs to reproduce binocular vision [19]. Despite the progress, these systems remain unsatisfactory due to inadequate visual cues such as dynamical cues [22]. Building upon this evidence, we maintain stereoscopic streaming from the eye camera as a high-fidelity primary display, and complement other visual cues via two wrist cameras. As shown in Fig. 3 (c), CaFe-TeleVision features two critical designs: on-demand display and situated visualization.

1) On-demand wrist-view display: Many multi-view VR applications non-stop render all views during runtime, regardless of whether a view is useful for user decisions and behaviors, as shown in Fig. 4. This "always-on" rendering strategy causes visual distraction and occlusion. Hence, we design an on-demand display strategy that allows operators to toggle the visibility of wrist-camera views using buttons.

A practical scenario is shown in Fig. 3 (c). When grasping a banana, the eye-camera view offers operators adequate visual information to locate and pick it without additional views. However, the other case is for a strawberry. Occlusion caused by gripper itself obscures successful grasping. In this circumstance, operators typically activate the right wrist-camera view via R4 to obtain close-up observations. This on-demand pattern improves depth perception and detail awareness, leading to a high grasping success rate and a low cognitive load.

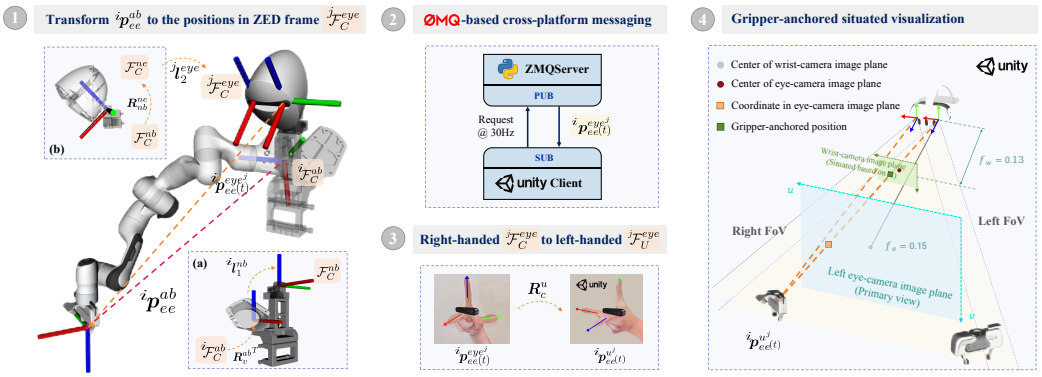


Fig. 5: Technical pipeline of gripper-anchored situated visualization. Specifically, step 1 calculates the gripper-to-eye translation. Step 2 messages the translation to the unity client via ZeroMQ. Step 3 transforms the translation based on right-handed rule to the one in left-handed rule. Step 4 situates the wrist-camera view near the gripper-anchored position.

2) Gripper-anchored situated visualization: Situated visualization [35] is an emerging technique that renders graphical elements within their spatially relevant contexts (i.e., situatedness). Compared with static layout, high situatedness is shown to support understanding and decision-making [34]. As such, a gripper-anchored situatedness is implemented in Unity, geometrically aligning the i-side wrist-camera view with i-side gripper in the eye-camera view. This gripper-anchored pattern is based on the observation that operators always cry for referring additional visual cues when performing contact-rich manipulation, with the gaze naturally focusing on the gripper. So the complementary cues are presented close to the gripper to reduce attention shift, thereby improving cognitive ergonomics in multi-view processing (see Fig. 4).

To preserve stereoscopic perception, two paired views are maintained to cast the observations from i-side wrist camera. Formally, let $j \in \{\mathrm{L},\mathrm{R}\}$ denote the eye side. Fig. 5 illustrates the four-step pipeline of how to situate the i-side wrist-camera view for j-side eye in the virtual Unity scene, exemplified with $i=R$ and $j=L$. First, the position of the i-side gripper in CURI's j-side eye-camera frame ${}^{j}\mathcal{F}_{C}^{eye}$ is:

$$ {}^{i}\boldsymbol{p}_{ee}^{eye^j}(t) = \boldsymbol{R}_{nb}^{ne}({}^{i}\boldsymbol{R}_{w}^{ab\,T}\,{}^{i}\boldsymbol{p}_{ee}^{ab}(t) + {}^{i}\boldsymbol{l}_{1}^{nb}) + {}^{j}\boldsymbol{l}_{2}^{eye}, \tag{7} $$

where $\boldsymbol{R}_{nb}^{ne}$ is the rotation matrix from neck base frame $\mathcal{F}_{C}^{nb}$ to neck end frame $\mathcal{F}_{C}^{ne}$, ${}^{i}\boldsymbol{l}_{1}^{nb}$ and ${}^{j}\boldsymbol{l}_{2}^{eye}$ represent constant translation vectors to frame $\mathcal{F}_{C}^{nb}$ and ${}^{j}\mathcal{F}_{C}^{eye}$, respectively. Second, ${}^{i}\boldsymbol{p}_{ee}^{eye^j}(t)$ is messaging to Unity via ZeroMQ at subscribing frequency 30 Hz. Third, the rotation transformation from real-world frame ${}^{j}\mathcal{F}_{C}^{eye}$ to frame ${}^{j}\mathcal{F}_{U}^{eye}$ in Unity is defined by $\boldsymbol{R}_{c}^{u}$, so ${}^{i}\boldsymbol{p}_{ee}^{u^j}(t) = \boldsymbol{R}_{c}^{u}\,{}^{i}\boldsymbol{p}_{ee}^{eye^j}(t)$. Fourth, to overlay the primary view (focal length f_e) in Unity, the focal length for the wrist-camera imaging plane is set to $f_w < f_e$. The i-side gripper-anchored position ■ for the wrist-camera imaging plane in j-side eye is projected as:

$$ {}^{i}\boldsymbol{p}_{anchor}^{u^j}(t) = \begin{bmatrix} {}^{i}p_{ee,x}^{u^j}(t)\, f_w / {}^{i}p_{ee,z}^{u^j}(t) \\ {}^{i}p_{ee,y}^{u^j}(t)\, f_w / {}^{i}p_{ee,z}^{u^j}(t) \\ f_w \end{bmatrix} \tag{8} $$

Finally, the i-side wrist-camera view for j-side eye is positioned on ●, with pre-defined in-plane translation and scale related to ■ to alleviate occlusion between views.

IV. Experiments

We conduct a user study to evaluate the performance of teleoperation systems. This section introduces the experimental setup, followed by evaluation metrics concerning success rate, efficiency, and ergonomics. Finally, we summarize quantitative and qualitative results, supported by statistical analysis.

A. Experimental Setup

1) Participant Profile: For user study, 24 participants (16 male and 8 female) were recruited through university-wide open invitation, aged between 19 and 33 years, with an average age of 24.96 and a standard deviation (SD) of 3.34. Their prior teleoperation experience and VR experience were assessed using a 5-point Likert scale questionnaire (1 for "Expert" and 5 for "Novice"), averaging scores of 3.2 for teleoperation (SD=1.58) and 3.4 for VR experience (SD=1.32), respectively.

2) Task Suite: To challenge the capabilities of teleoperation systems, we elaborately designed six tasks, including *insert torus*, *grasp fruits*, *pour tea*, *twist cap*, *hang towel*, *pack bag*. Fig. 6 shows critical phases and the characteristics for each task. Specifically, challenges involve: fine-grained orientation control, handling a large range of rotation outside of the ergonomic joint range of operators, and visual occlusion. All characteristics set an effective testbed to evaluate the strengths of our system. More details are available in the supplementary video (e.g., the prescribed time limit for each task).

3) Baseline Systems: To evaluate the effectiveness of CaFe-TeleVision, we compare it with three representative systems. These baselines adopt standard retargeting and perception schemes, typical in state-of-the-art teleoperation systems. For brevity, we refer to N as natural mode, R as relative mode, S as stereo visual feedback, and SL as static multi-view layout.

- **N-S**: Adopt natural mode for retargeting and only streams the stereo video from the eye camera, without the wrist-camera views, similar to Bunny-VisonPro [8].
- **CaFe-SL**: A variant of CaFe-TeleVision that renders multiple views in the static layout instead of on-demand situated visualization technique.
- **R-TeleVision**: A variant of CaFe-TeleVision that replaces coarse-to-fine mechanism with relative mode for retargeting.

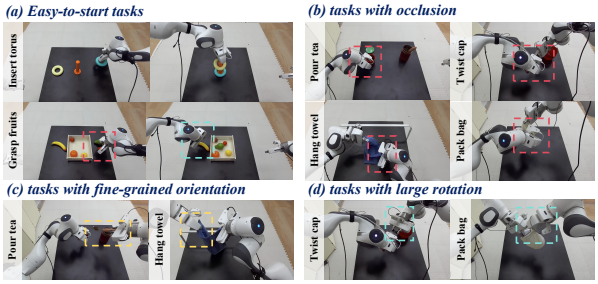


Fig. 6: Illustration of key phases and characteristics of six manipulation tasks. (a) shows two easy-to-start tasks for novices. (b)-(d) denotes tasks facing challenges in occlusion, fine-grained orientation control, and large rotation, respectively. Each challenge is highlighted with a distinct colored box.

Each time, retargeting is toggled by holding L2/R2 button, and the relative pose transformation depends on the initial pose when pressing the buttons.

4) Experimental Design: Pilot studies indicate that the difficulty of several tasks (such as *pour tea* and *pack bag*) exceeds what novices can successfully complete within a short practice period (e.g., 5 minutes). Incorporating excess low-quality results overwhelms the statistical reliability of system comparison. Towards this, the experiments are designed into two parts. First, two easy-to-start tasks (i.e., *insert torus* and *grasp fruits*) are selected for all participants to perform. The results from this wide demographic enhance the comparison analysis with high statistical power. Second, three expert participants evaluate all six tasks, where experts denote those who self-rated both teleoperation and VR experience as 1 (indicating "Expert"). Benefiting from low proficiency-related variability, failure case analysis on the results is thus expected to reveal the inherent capabilities and limitations of each system in handling challenging scenarios. The experimental durations for each non-expert and expert are within 3 hours and 12 hours, respectively. The compensation depends on the working hours, at a rate of HK $64 per hour.

All experiments employ a within-subjects design, where each participant evaluates all four systems across testing tasks, performing five trials with each system per task to enhance measurement reliability. In total, each participant executes 20 trials per task. To alleviate learning and order effects, we implemented rigorous order control throughout the experiments. Specifically, a partial counterbalancing design with full permutations of systems was adopted for the all-participant experiment. In detail, a complete counterbalancing of task order was employed, evenly assigning two task order sequences to all participants. We then fully covered all possible system permutations among four systems and assigned a unique system order to each participant. For the expert experiment, six tasks are grouped by bimanual coordination levels (see Fig. 1), and the group order assigned to each expert is based on Latin Square counterbalancing. The task order within groups and the system order were randomized.

5) Experimental Procedure: Participants received an introduction to teleoperation and experimental objectives. The task and system sequence for evaluation were then assigned based on the established order control. Following this, they calibrated the Xsens motion tracking system based on provided instructions and underwent a 10-minute training session to familiarize themselves with operational keypoints of these systems. For each testing task, they were given a 5-minute practice period before officially recording their performance. After that, the quantitative results were measured during the experiments. At the end, participants completed relevant questionnaires to assess their qualitative subjective feelings.

TABLE II: Quantitative teleoperation performance on *insert torus* and *grasp fruits* task, in terms of *success rate (SR)* and *completion time (Time)* in seconds. The results are average over 120 trials from all participants. Note: *TeleVision* is abbreviated as *Tele*.

Method	Insert torus		Grasp fruits	
	SR ↑	Time ↓	SR ↑	Time ↓
N-S	96.81%±17.67%	36.97±12.08	70.00%±46.08%	79.99±35.70
CaFe-SL	94.68%±22.56%	39.52±14.84	73.33%±41.47%	80.83±37.03
R-Tele	81.92%±38.40%	49.04±18.36	54.44%±50.08%	100.79±39.19
CaFe-Tele	**97.87%±14.51%**	**33.90±11.29**	**78.89%±41.04%**	**72.44±34.89**

TABLE III: Post-hoc Nemenyi tests of *completion time* metric on *insert torus* and *grasp fruits* tasks. Significant pairs ($p<0.05$) are listed with p-values to six decimal places.

Task	Significant pair	*p*-value
Insert torus	CaFe-TeleVision vs. R-TeleVision	0.000003
	CaFe-TeleVision vs. N-S	0.020126
Grasp fruits	CaFe-TeleVision vs. R-TeleVision	0.000002
	CaFe-SL vs. R-TeleVision	0.020126

B. Evaluation Metrics

1) Quantitative Measurements: To quantitatively analyze performance for specific tasks, two metrics are measured during the experiments. (a) *Success Rate (SR)*: the ratio of successful trials to total trials; (b) *Completion Time (Time)*: the average task completion time across all trials. A task trial is deemed a failure under the following conditions. First, it exceeds the prescribed time. Second, for fine-grained tasks, if the tea is split or towel is not hung flat, the trial is also counted as a failure. Trials above twice the time limit are terminated.

2) Qualitative Assessments: To qualitatively assess the subjective feelings regarding system workload and acceptance, two typical questionnaires were conducted after the experiments. (a) *NASA-TLX*: a six-dimensional task load index, with higher values reflecting greater perceived difficulty in experiments. (b) *System Usability Scale (SUS)*: a 10-item instrument designed to assess subjective usability, with higher values showing a stronger user acceptance of the method.

C. All-Participant Performance Analysis

Table II presents the quantitative results on two easy-to-start tasks: *insert torus* and *grasp fruits*. Each item is calculated over 120 trials from all participants, ensuring measurement reliability. Overall, CaFe-Television achieves superior performance, especially in *grasp fruit* task. To evaluate systems statistically, Friedman tests with post-hoc Nemenyi tests were conducted on *completion time* metric separately for each task. Friedman tests confirmed significant differences among four systems, with $\chi^2(3)=25.30$, $p=0.000013$ for *insert torus* and $\chi^2(3)=30.11$,

TABLE IV: Statistical analysis on six-dimensional NASA-TLX scores from all participants, using Friedman tests with post-hoc Nemenyi tests for pairwise evaluation. Assessments include mental demand (MD), physical demand (PD), temporal demand (TD), performance, effort, and frustration. P-values are reported to four decimal places, with significance ($p<0.05$) bolded.

		Six-dimensional task load index						Mean
		MD	PD	TD	Performance	Effort	Frustration	
Friedman test	$\chi^2(3)$	11.00	16.16	15.19	14.49	14.20	13.20	17.93
	p-value	**0.0117**	**0.0010**	**0.0012**	**0.0023**	**0.0026**	**0.0042**	**0.0005**
Post-hoc system vs. system pair	CaFe-TeleVision vs. N-S	0.3709	**0.0037**	**0.0179**	**0.0015**	**0.0305**	**0.9999**	**0.0024**
	CaFe-TeleVision vs. CaFe-SL	0.0911	0.6675	0.1371	0.3064	0.0789	0.8903	**0.0215**
	CaFe-TeleVision vs. R-TeleVision	**0.0149**	0.9952	**0.0024**	0.1762	**0.0069**	0.0911	**0.0015**
	N-S vs. CaFe-SL	0.8903	0.1049	0.8650	0.2227	0.9842	0.8650	0.9130
	N-S vs. R-TeleVision	0.5154	**0.0083**	0.9328	0.3709	0.9640	0.1049	0.9994
	CaFe-SL vs. R-TeleVision	0.9130	0.8067	0.5154	0.9907	0.8370	**0.0124**	**0.8650**

TABLE V: Quantitative teleoperation performance from expert participants on six challenging tasks, in terms of *success rate (SR)* and *completion time (Time)* in seconds.

Method	Insert torus		Grasp fruits		Pour tea		Twist cap		Hang towel		Pack bag	
	SR ↑	Time ↓	SR ↑	Time ↓	SR ↑	Time ↓	SR ↑	Time ↓	SR ↑	Time ↓	SR ↑	Time ↓
N-S	**100%**	30.77	**100%**	57.59	46.67%	60.38	73.33%	117.02	73.33%	51.67	33.33%	145.36
CaFe-SL	**100%**	**28.01**	**100%**	52.52	86.67%	66.49	100%	100.55	86.67%	54.64	100%	110.20
R-TeleVision	**100%**	36.42	**100%**	70.11	80%	68.81	53.33%	142.49	100%	57.55	66.67%	132.52
CaFe-TeleVision	**100%**	29.31	**100%**	**46.25**	**100%**	**57.74**	**100%**	**86.53**	**100%**	**48.89**	**100%**	**103.01**

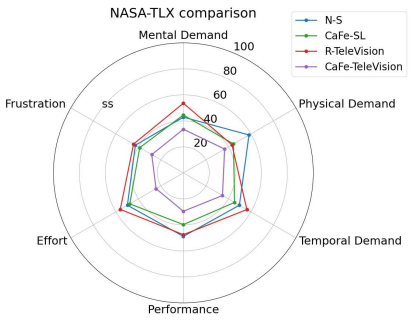
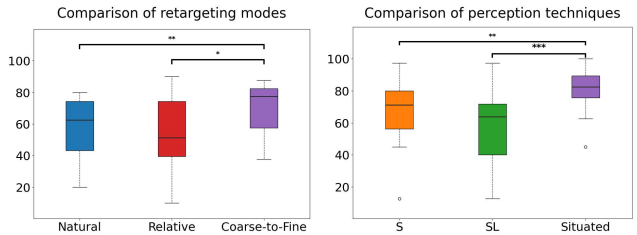


Fig. 7: NASA-TLX results from all participants on six-dimensional task load scores for four systems. Our system achieves the minimal enclosed area, validating a lower physical and mental workload.



(a) Retargeting modes (b) Perception techniques

Fig. 8: System usability scale (SUS) results from all participants, illustrating higher acceptance of our proposed schemes. (a) compares retargeting modes: natural mode, relative mode, and our coarse-to-fine mechanism. (b) contrasts perception techniques: stereo streaming (S), static multi-view layout (SL), and our on-demand situated visualization (Situated). Significant pairwise differences are mark by asterisks (*: $p<0.05$, **: $p<0.01$, ***: $p<0.001$).

$p=0.000001$ for *grasp fruits*. Post-hoc Nemenyi tests revealed specific significant pairs, as reported in Table III. We can see: CaFe-TeleVision completion times were significantly lower than those of R-TeleVision ($p < 0.05$), which demonstrates that our system supports more efficient operation compared to relative mode. For *grasp fruits* task, our system established statistical significance with N-S, validating the effectiveness of the coarse-to-fine control mechanism in the scenario requiring large wrist rotation. But no significant difference was found between CaFe-SL and N-S, indicating the control advantage brought by this mechanism is diminished by the side effects from static multi-view perception. This observation thus reflects that our on-demand situated visualization is desired for multi-view processing.

Fig. 7 illustrates the six-dimensional workload assessment via NASA-TLX for four systems, with statistical analysis results shown in Table IV. The radar chart demonstrates that CaFe-TeleVision achieves the minimal enclosed area, which confirms that the proposed mechanisms successfully reduce the task load ($p<0.05$, Friedman test). Post-hoc Nemenyi tests report pairwise significances between systems. Notably, the improvement is particularly pronounced in *Effort* and *Performance* dimensions compared to baseline systems, with average improvements of 16.87% and 15.83%, respectively. These findings indicate enhanced ergonomics by CaFe-TeleVision, thereby boosting overall performance.

Fig. 8 (a) and (b) present the System Usability Scale (SUS) results. Friedman tests indicated significant differences among retargeting modes ($\chi^2(3)=11.37$, $p=0.003393$) and perception techniques ($\chi^2(3) = 22.07$, $p = 0.000016$). Post-hoc Nemenyi tests confirmed pairwise significance, highlighting higher acceptance of our proposed schemes. Statistical evidences show that our coarse-to-fine retargeting was significantly preferred over standard modes ($p < 0.05$), especially over natural mode ($p < 0.01$). Benefiting from its intuitiveness and seamless switching characteristics, our mechanism jointly optimizes efficiency and physical ergonomics, surpassing relative mode by an average of 25.51%. For perceptual feedback, on-demand situated visualization outperformed static multi-view layouts

($p < 0.001$), affirming the cognitive ergonomics benefits of integrating situated visualization techniques in teleoperation. Additionally, we note a higher acceptance (averaging 20.89%) of single-view stereo streams over static multi-view displays. We attribute this preference to the detrimental effects of static layouts, such as visual distraction and occlusion, which overshadow the advantages of augmented visual cues during the experiments. This further validates the effectiveness of our perception scheme.

D. Expert Performance Analysis on Challenging Tasks

Table V reports the quantitative results from three expert participants on six bimanual teleoperation tasks, facing challenges in orientation and occlusion handling. On average, CaFe-TeleVision consistently improves performance, boosting success rates by up to 28.89% and efficiency by 26.81%.

Owing to low proficiency-related variability in expert-derived results, we can identify the inherent strengths and limitations of each system through failure analysis. For fine-grained tasks, N-S failed in 8 trials on *pour tea* and 4 trials on *hang towel* due to difficulties in fine adjustment under natural mapping mode. In *twist cap* and *pack bag* tasks, operators also struggled to exert large wrist rotation under natural mode due to anatomical limits, leading to uncomfortable postures and timeouts. CaFe-SL failed twice in *pour tea* and *hang towel* tasks, as the fixed wrist views partially occluded operational areas, slowing teleoperation and thus causing timeout failure. R-TeleVision suffers from low efficiency due to the inevitable pauses and non-intuitive orientation retargeting in relative mode, ultimately resulting in severe timeout failures. In contrast, CaFe-TeleVision overcomes these limitations with its coarse-to-fine retargeting and situated visualization technique, leading to superior performance across all complex tasks.

V. Conclusion

We proposed CaFe-TeleVision, a coarse-to-fine teleoperation system with immersive situated visualization for enhanced ergonomics. Comprehensive experiments are conducted in the user study, confirming its superior performance and enhanced ergonomics with statistical significance. Moreover, our method bridges teleoperation challenges in a practical and deployable manner, which can be integrated as plug-and-play modules into current VR-based teleoperation systems, yielding immediate and substantial performance improvements.

Limitations and Future Work. First, the transparent spatial mapping in our joystick-assisted mode assumes the gripper approach direction approximately aligns with the viewpoint, which may not generalize to arbitrary configurations. Future work will explore viewpoint-dependent spatial mapping. Second, integration of additional feedback modalities, such as haptic feedback, is desired to enhance cognitive ergonomics, especially in tactile-intensive tasks. Third, developing a collision avoidance mechanism in teleoperation systems is important for ensuring safety.

References

[1] T. B. Sheridan, "Telerobotics," *Automatica*, vol. 25, no. 4, pp. 487–507, 1989.

[2] Z. Tang *et al.*, "Csgp: Closed-loop safe grasp planning via attention-based deep reinforcement learning from demonstrations," *IEEE RA-L*, vol. 8, no. 6, pp. 3158–3165, 2023.

[3] C. Chi *et al.*, "Diffusion policy: Visuomotor policy learning via action diffusion," *IJRR*, vol. 44, no. 10-11, pp. 1684–1704, 2023.

[4] M. J. Kim *et al.*, "Openvla: An open-source vision-language-action model," *CoRL*, 2025.

[5] Q. Rouxel *et al.*, "Extremum flow matching for offline goal conditioned reinforcement learning," in *IEEE-RAS Humanoids*, 2025.

[6] M. Schwarz *et al.*, "Robust immersive telepresence and mobile telemanipulation: Nimbro wins ana avatar xprize finals," in *IEEE-RAS Humanoids*, 2023.

[7] T. Z. Zhao *et al.*, "Learning fine-grained bimanual manipulation with low-cost hardware," *RSS*, 2023.

[8] R. Ding *et al.*, "Bunny-visionpro: Real-time bimanual dexterous teleoperation for imitation learning," *CoRR*, 2024.

[9] Y. Liu *et al.*, "Avr: Active vision-driven robotic precision manipulation with viewpoint and focal length optimization," *arXiv preprint arXiv:2503.01439*, 2025.

[10] K. Darvish *et al.*, "Teleoperation of humanoid robots: A survey," *IEEE Transactions on Robotics*, vol. 39, no. 3, pp. 1706–1727, 2023.

[11] Q. Rouxel *et al.*, "Multicontact motion retargeting using whole-body optimization of full kinematics and sequential force equilibrium," *IEEE/ASME TMECH*, vol. 27, no. 5, pp. 4188–4198, 2022.

[12] A. Mandlekar *et al.*, "Roboturk: A crowdsourcing platform for robotic skill learning through imitation," in *CoRL*, 2018.

[13] I. Ozdamar *et al.*, "A shared autonomy reconfigurable control framework for telemanipulation of multi-arm systems," *IEEE RA-L*, vol. 7, no. 4, pp. 9937–9944, 2022.

[14] T. Lin *et al.*, "Learning visuotactile skills with two multifingered hands," *IEEE ICRA*, 2024.

[15] W.-k. Fung *et al.*, "A case study of 3d stereoscopic vs. 2d monoscopic tele-reality in real-time dexterous teleoperation," in *IEEE/RSJ IROS*, 2005.

[16] S. Livatino *et al.*, "Stereo viewing and virtual reality technologies in mobile robot teleguide," *IEEE Trans. on Robotics*, vol. 25, no. 6, pp. 1343–1355, 2009.

[17] Y. Ishiguro *et al.*, "High speed whole body dynamic motion experiment with real time master-slave humanoid robot system," in *IEEE ICRA*, 2018.

[18] L. Penco *et al.*, "A multimode teleoperation framework for humanoid loco-manipulation: An application for the icub robot," *IEEE RAM*, vol. 26, no. 4, pp. 73–82, 2019.

[19] X. Cheng *et al.*, "Open-television: Teleoperation with immersive active visual feedback," *CoRL*, 2024.

[20] J. Ryu *et al.*, "Functional ranges of motion of the wrist joint," *The Journal of hand surgery*, vol. 16, no. 3, pp. 409–419, 1991.

[21] R. K. Mehta, "Integrating physical and cognitive ergonomics," *IISE Trans. Occup. Ergon. Hum. Factors*, vol. 4, no. 2-3, pp. 83–87, 2016.

[22] F. El Jamiy *et al.*, "Survey on depth perception in head mounted displays: distance estimation in virtual reality, augmented reality, and mixed reality," *IET Image Processing*, vol. 13, no. 5, pp. 707–712, 2019.

[23] N. Bressa *et al.*, "What's the situation with situated visualization? a survey and perspectives on situatedness," *IEEE TVCG*, vol. 28, no. 1, pp. 107–117, 2021.

[24] G. H. Martınez, "Openpose: Whole-body pose estimation," *Ph.D. thesis*, 2019.

[25] D. Roetenberg *et al.*, "Xsens mvn: Full 6dof human motion tracking using miniature inertial sensors," *Xsens Motion Technologies BV, Tech. Rep*, 2009.

[26] F. P. Audonnet *et al.*, "Immertwin: A mixed reality framework for enhanced robotic arm teleoperation," *arXiv preprint arXiv:2409.08964*, 2024.

[27] V. Patil *et al.*, "Radiance fields for robotic teleoperation," in *IEEE/RSJ IROS*, 2024.

[28] M. Laghi *et al.*, "Shared-autonomy control for intuitive bimanual tele-manipulation," in *IEEE-RAS Humanoids*, 2018.

[29] A. Iyer *et al.*, "Open teach: A versatile teleoperation system for robotic manipulation," *CoRL*, 2025.

[30] R. Wen *et al.*, "Collaborative bimanual manipulation using optimal motion adaptation and interaction control: Retargeting human commands to feasible robot control references," *IEEE RAM*, vol. 31, no. 4, pp. 68–80, 2023.

[31] A. Ajoudani *et al.*, "Tele-impedance: Teleoperation with impedance regulation using a body–machine interface," *IJRR*, vol. 31, no. 13, pp. 1642–1656, 2012.

[32] S. Yang *et al.*, "Ace: A cross-platform visual-exoskeletons system for low-cost dexterous teleoperation," *CoRL*, 2025.

[33] O. Kanoun *et al.*, "Kinematic control of redundant manipulators: Generalizing the task-priority framework to inequality task," *IEEE T-RO*, 2011.

[34] Z. Wen *et al.*, "Effects of view layout on situated analytics for multiple-view representations in immersive visualization," *IEEE TVCG*, vol. 29, no. 1, pp. 440–450, 2022.

[35] S. M. White, *Interaction and presentation techniques for situated visualization*. Columbia University, 2009.

[36] Y. Chen *et al.*, "Unified model predictive interaction control integrating impedance matching and constraint optimization," in *IEEE ICCA*, 2025.