frontiers | Frontiers in Robotics and AI

TYPE Original Research
PUBLISHED 05 February 2026
DOI 10.3389/frobt.2026.1749105

OPEN ACCESS

EDITED BY
Tao Teng,
The Chinese University of Hong Kong, China

REVIEWED BY
Junling Fu,
Polytechnic University of Milan, Italy
Jianzhuang Zhao,
Italian Institute of Technology (IIT), Italy

*CORRESPONDENCE
Luka Peternel,
✉ l.peternel@tudelft.nl

RECEIVED 18 November 2025
REVISED 08 January 2026
ACCEPTED 14 January 2026
PUBLISHED 05 February 2026

CITATION
Jekel HHA, Díaz Rosales A and Peternel L (2026) Visio-verbal teleimpedance interface: enabling semi-autonomous control of physical interaction via eye tracking and speech.
Front. Robot. AI 13:1749105.
doi: 10.3389/frobt.2026.1749105

Visio-verbal teleimpedance interface: enabling semi-autonomous control of physical interaction via eye tracking and speech

Henk H. A. Jekel[1], Alejandro Díaz Rosales[1,2] and Luka Peternel[1]*

[1]Department of Cognitive Robotics, Delft University of Technology, Delft, Netherlands, [2]European Organization for Nuclear Research (CERN), Meyrin, Switzerland

The paper presents a visio-verbal teleimpedance interface for commanding 3D stiffness ellipsoids to the remote robot with a combination of the operator's gaze and verbal interaction. The gaze is detected by an eye-tracker, allowing the system to understand the context in terms of what the operator is currently looking at in the scene. Along with verbal interaction, a Vision-Language Model (VLM) processes this information, enabling the operator to communicate their intended action or provide corrections. Based on these inputs, the interface can then generate appropriate stiffness matrices for different physical interaction actions. To validate the proposed visio-verbal teleimpedance interface, we conducted a series of experiments on a setup including a Force Dimension Sigma.7 haptic device to control the motion of the remote Kuka LBR iiwa robotic arm. The human operator's gaze is tracked by Tobii Pro Glasses 2, while human verbal commands are processed by a VLM using GPT-4o. The first experiment explored the optimal prompt configuration for the interface. The second and third experiments demonstrated different functionalities of the interface on a slide-in-the-groove task.

KEYWORDS

gaze tracking, impedance control, teleoperation, verbal interaction, vision-language model

1 Introduction

Teleoperation is a key technology enabling remote human control and teaching of robots in scenarios such as disaster response, robot-assisted surgery, inspection & maintenance, space exploration, and hazardous environment operations (Si et al., 2021). While autonomous robots excel in structured environments like manufacturing, they struggle to adapt to dynamic, unstructured conditions due to limited cognitive flexibility. In that respect, teleoperation offers more adaptability by integrating humans into the robot control loop, allowing operators to issue motion commands through interfaces such as haptic devices, joysticks, and motion capture systems.

Nevertheless, controlling the motion alone makes the execution of complex tasks in interaction with unstructured and unpredictable environments difficult (Suomalainen et al., 2022). Impedance control enables the control of the relationship between forces and motion and simplifies the execution of tasks with physical interactions

(Naceri et al., 2021). In this case, stiffness is the most important parameter of impedance as it controls how soft or stiff the robot is when interacting with fragile objects. Teleimpedance is a concept that allows human operators to control the stiffness of the remote robot through various interfaces (Peternel and Ajoudani, 2022).

Existing teleimpedance command interfaces can be broadly categorised into manual impedance control and automated impedance control, depending on whether the human or the robot automation determines the impedance of the interaction (Peternel and Ajoudani, 2022). Manual control approaches allow the human operator to directly adjust stiffness through different inputs. One common method involves estimating the operator's stiffness with muscle activity using electromyography (EMG), which is then mapped to the robot's impedance (Ajoudani et al., 2012; Yang et al., 2018; Park et al., 2018; Buscaglione et al., 2022). While EMG-based teleimpedance interfaces enable intuitive multi-degree of freedom (DoF) control of impedance, they require sensor placement and calibration procedures and have limited generalisability. Simplification in terms of the number of EMG sensors alleviates some of these issues; however, they require an additional motion capture system to retain some of the multi-DoF functionality (Ajoudani et al., 2018). Interfaces based on biosignals are also subject to a neuromechanical coupling effect between the commanded stiffness and force feedback, which takes away some of the control due to reflexes (Doornebosch et al., 2021).

More practical teleimpedance interfaces are based on controlling stiffness with force grip sensor (Walker et al., 2011) and buttons (Garate et al., 2021), joysticks/scroll-wheels (Kraakman and Peternel, 2024), perturbations on haptic device (Gourmelen et al., 2021), touchscreens (Peternel et al., 2021), or augmented-reality (Díaz Rosales et al., 2024). Nevertheless, they either have limited control over the stiffness ellipsoid in terms of DoF, perturb the operator, or take the operator's attention away from the task and impose a high cognitive workload on the operator.

In contrast to manual impedance control, automated impedance control systems remove the operator from the impedance control loop, allowing the robot autonomy to determine the appropriate stiffness. For example, in the method in (Michel et al., 2021), the robot used torque sensors to measure physical interaction forces and autonomously adjust stiffness for task stabilisation, ensuring compliance when needed and increasing rigidity during disturbances. Similarly, the work in (Siegemund et al., 2024) proposed a vision-based system where the robot detects material properties and geometry of an object that the operator was about to touch with the remote robot. Based on that, the robot autonomy preemptively sets optimal stiffness values for the expected physical interaction. While these autonomous approaches improve safety and reduce operator workload, they remove direct human input and reduce the operator's ability to intervene.

Despite advancements in teleimpedance interfaces, current methods either require continuous manual adjustments from the operator, which are cognitively demanding or rely on automation, which limits human input. To bridge this gap, this work introduces a novel visio-verbal teleimpedance interface that follows the shared control paradigm (see Figure 1). Gaze-based interfaces have been commonly applied to control robot manipulation on a higher level, e.g., to infer the goal while the robot autonomously controls motion toward that goal (Admoni and Srinivasa, 2016; Aronson et al., 2021). Nevertheless, this does not allow the human control over low-level actions such as motion and impedance, which can often be too complex for robots to handle autonomously in difficult physical interaction. The method we propose here enables the operator to have low-level control as well as augmentation of high-level control with the gaze and speech interfaces. Table 1 provides a conceptual comparison of state-of-the-art methods.

By leveraging gaze and speech, two natural and intuitive communication modalities, the proposed interface allows operators to configure the robot's 3D stiffness ellipsoid, without diverting their visual attention from the task. The operators can inform the robot autonomy what they intend to do by conversation and gaze, and if the decisions of autonomy are unsatisfactory, corrections can be communicated. The conversation could be conducted by asking the operator to write down the request. However, during a study in (Díaz Rosales et al., 2024), it was noted that the expert operator participant indicated a preference to control the stiffness using voice commands to keep their hands on the robot's movement controls. As a result, speech recognition modality was explored for the interface proposed in this paper.

Unlike previous methods, which either require direct manual adjustments or rely entirely on automation, this approach distributes cognitive workload between the human operator and a Vision-Language Model (VLM). The interface provides gaze-based contextual awareness and interprets verbal commands to dynamically generate a stiffness matrix that optimises the robot's interaction with its environment. One of the key contributions of this work is to provide the first teleimpedance interface that combines gaze and speech modalities.

The rapid advancements in Deep Learning, particularly the transformer architecture (Vaswani et al., 2017) and the increasingly larger model sizes (Kaplan et al., 2020) have enabled breakthrough applications in speech-to-text, text-to-speech, and natural language processing combined with computer vision through VLMs. A popular example is ChatGPT, which uses a VLM to process both textual and visual inputs. VLMs have been recently applied to augment teleoperation for robot manipulation tasks (Fu et al., 2025; Cui et al., 2025; Liu et al., 2025). Nevertheless, these methods do not enable impedance control, which is crucial for complex physical interaction, and do not leverage real-time verbal or gaze inputs, which can offer rapid natural user interaction. While standard VLMs can generate highly contextual responses, they typically lack direct awareness of user focus unless extensively prompted. Recently, researchers proposed GazeGPT, which incorporates mobile eye-tracking to enhance context awareness by identifying where a user is looking (Konrad et al., 2024). Building upon these advancements, we develop VLM-driven teleimpedance control integrating eye tracking into a visio-verbal system.

In the experimental work, we first conducted an extensive prompt parameter optimisation study, specific to the use of VLM in teleimpedance, where we investigated and compared the effects of different settings (system role, amount of priors, and image resolution) on the stiffness matrix predictions. To demonstrate and validate the key functionalities of the developed novel visio-verbal teleimpedance interface, we conducted experiments on a setup where Force Dimension Sigma.7 haptic device controls the motion of the remote Kuka LBR iiwa robotic arm. The human operator's gaze is tracked by Tobii Pro Glasses 2, while human

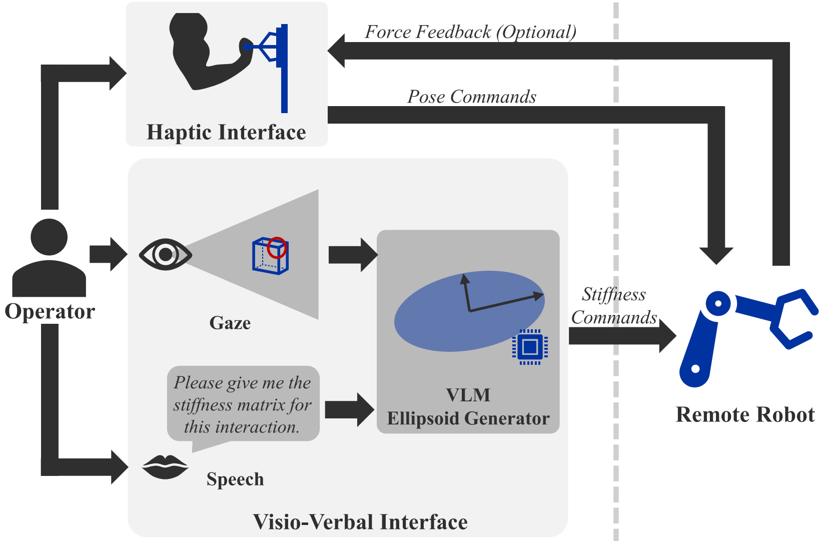


FIGURE 1
Diagram presenting the visio-verbal teleimpedance interface inside a teleimpedance setup. The local and remote sites are separated by the dotted line. At the local site, the operator controls the remote robot using a haptic device to command motion. The visio-verbal interface operates in parallel, allowing the operator to configure the desired stiffness using gaze and speech inputs. This stiffness configuration, along with motion commands, is then transmitted to the remote robot.

TABLE 1 Conceptual comparison of state-of-the-art method specific for teleoperated manipulation using key elements of this work: teleimpedance, gaze, and VLM. Colours indicate general preference for each quality, where green is desired, and red is not desired. Impedance control and gaze input signify whether the methods enable those functionalities. Regarding human control, high means the operator has full control over remote robot actions, medium means the human can control low-level motion but impedance control is fully relegated to the robot, while low means that the operator can only give high-level goals and/or there is no variable impedance control. Robot autonomy means how autonomous the robot can/should be. Regarding calibration overhead for VLM and visio-verbal interfaces, it is assumed that model training is included as part of calibration overhead.

Interface type	Impedance Control	Gaze Input	Human Workload	Human Control	Robot Autonomy	Calibration Overhead	Required Sensors
Muscle Ajoudani et al. (2018)	Yes	No	Medium	High	Low	High	EMG & motion capture
Virtual Peternel et al. (2021) Díaz Rosales et al. (2024)	Yes	No	High	High	Low	Low	Touchscreen or AR headset
Haptic Michel et al. (2021)	Yes	No	Low	Medium	Medium	Low	Force/torque sensor
Vision Siegemund et al. (2024)	Yes	No	Low	Medium	Medium	Low	Camera
Gaze Aronson et al. (2021)	No	Yes	Low	Low	High	Medium	Eye-tracking
VLM Fu et al. (2025)	No	No	Low	Low	High	High	Camera
Visio-verbal (proposed)	Yes	Yes	Low	Medium	High	High	Camera, mic & eye-tracking

verbal commands are processed by a VLM using GPT-4o. Two scenarios were investigated: the streamlined verbal mode of the interface, demonstrating fluid task execution, and the full mode, demonstrating the combined gaze and conversational aspects. The main contributions of this paper are: 1) the first-ever visio-verbal teleimpedance interface combining VLMs and gaze tracking, and 2) new insights about the effects of different settings on the stiffness matrix predictions in teleimpedance.

2 Methods

The concept of the proposed approach is illustrated by Figure 1, where gaze and speech are used to determine the appropriate stiffness configuration for the remote robot. This is done in a natural manner without diverting their visual attention from the teleoperation task. Through the operator's speech and gaze inputs, the VLM ellipsoid generator defines the stiffness ellipsoid for the given physical interaction. The visio-verbal teleimpedance interface was designed to enable hands-free adjustment of the orientation and shape of the 3D translational stiffness ellipsoid. This means that the human operator can use their hands exclusively to control the motion of the remote robot in real-time through a haptic device.

While we use a haptic device to command the motion, one could instead use alternative interfaces for motion commanding. For example, a depth camera can detect operator motion and is cheaper and less complex, but it is less precise and cannot provide force feedback. A haptic device can measure human arm motion precisely via encoders and can provide feedback about the forces that are felt by the remote robot, which can help with telepresence. The force feedback can be optionally turned off if there may be a delicate operation where force feedback-related instabilities might cause issues.

To achieve the stiffness generation, the interface captures a snapshot of the teleoperation scene after the operator says "capture", overlaying a red circle to indicate the operator's gaze estimate. Afterwards, the system processes the operator's verbal commands regarding their intention. The verbal interactions are processed through a speech-to-text module. The VLM then interprets these multimodal inputs to generate an updated stiffness matrix, which is applied directly to the remote robot. The advantage of additional gaze input to the VLM is to reduce the amount of speech needed to indicate the operator's intention. To facilitate verification, the system provides immediate feedback through both verbal confirmation and real-time visualisation of the updated stiffness ellipsoid, ensuring that adjustments align with the operator's intent.

2.1 Design requirements

Based on the discussed aspects in the introduction and analysis of related work (Peternel and Ajoudani, 2022), we formulated a set of design requirements to guide the development of the novel *visio-verbal* teleimpedance command interface. These requirements aim to address existing challenges and prioritise a human-centric design. The requirements are as follows:

1. Combine gaze tracking with verbal interaction for robot stiffness matrix generation.
2. Enable control of full 3D stiffness ellipsoid.
3. Minimise visual distractions to ensure the operator's uninterrupted focus on the task.
4. Minimal setting up and calibration procedures.
5. Avoid the neuromechanical coupling effect between the commended stiffness and force feedback.

To incorporate R1, the system follows a human-in-the-loop shared control paradigm, where the VLM assists in shaping and directing the stiffness ellipsoid, while still allowing the operator to refine its final configuration via voice commands. The system integrates a VLM to interpret high-level semantic inputs and gaze-based contextual awareness to quickly interpret what the operator is talking about in the remote scene. This provides a good middle ground between manual and autonomous variable impedance control. Using interpreted context from the speech and gaze, the system can generate a 3D stiffness ellipsoid to fit the given interaction in a specific task phase, thus achieving R2.

The combination of gaze and speech inputs with real-time multimodal processing simplifies stiffness modulation by eliminating the need for physical input devices. For instance, the teleimpedance interface that forms ellipsoids on a touchscreen allows for the generation of a 3D stiffness ellipsoid. However, this approach can distract the operator from the task at hand. In contrast, the proposed interface enables the operator to maintain focus on the task while an eye tracker monitors their gaze in the remote scene, thus achieving R3.

Since the microphone and eye tracker hardware are easy to use, the setup procedure is very short. Similarly, the calibration for the eye tracker takes only a minute. Therefore, avoiding the use of biosignals, we address R4. Finally, since the voice and gaze are both non-physical variables, there is no neuromechanical coupling effect between the commanded stiffness and force feedback that would result in a temporary loss of control over stiffness. This addresses R5.

2.2 Visio-verbal teleimpedance interface architecture

The primary objective in designing the architecture of the proposed visio-verbal teleimpedance interface is to ensure the seamless integration of all its components while maintaining modularity and interchangeability. This flexibility allows the system to adapt easily to different hardware configurations and future updates. To accomplish this, we adopted a container-based approach, allowing the software modules to operate in isolated environments, whether remotely or locally, while preventing version conflicts. This design minimises the setup time and technical overhead, in line with the requirement for minimal configuration (R4). Figure 2 shows all the containers used in this interface and the connections between them.

Some of these containers are responsible for controlling the devices that the operator interacts with, such as the eye tracker, the haptic device or the graphical user interface (GUI). Each device is managed by a different container. The GUI is used for calibration, image capture, audio recording, and device status. It serves as the first point of entry for the operator. All requests from the GUI are sent to the *Central Backend*, which processes and forwards them to the appropriate destination. The *Central Backend* manages all the information flow, taking the information from all the services.

On one side, we have the *Eye Tracker* container that is responsible for connecting to the eye-tracking device, whether through a cable or wirelessly. It manages the short initial eye-tracking calibration process, performs gaze estimation, and overlays the estimated gaze point onto a snapshot of the operator's view. As a private server, there is also the *Stiffness Database* that stores and provides generated stiffness matrices and ellipsoids, making them accessible to the *Central Backend*.

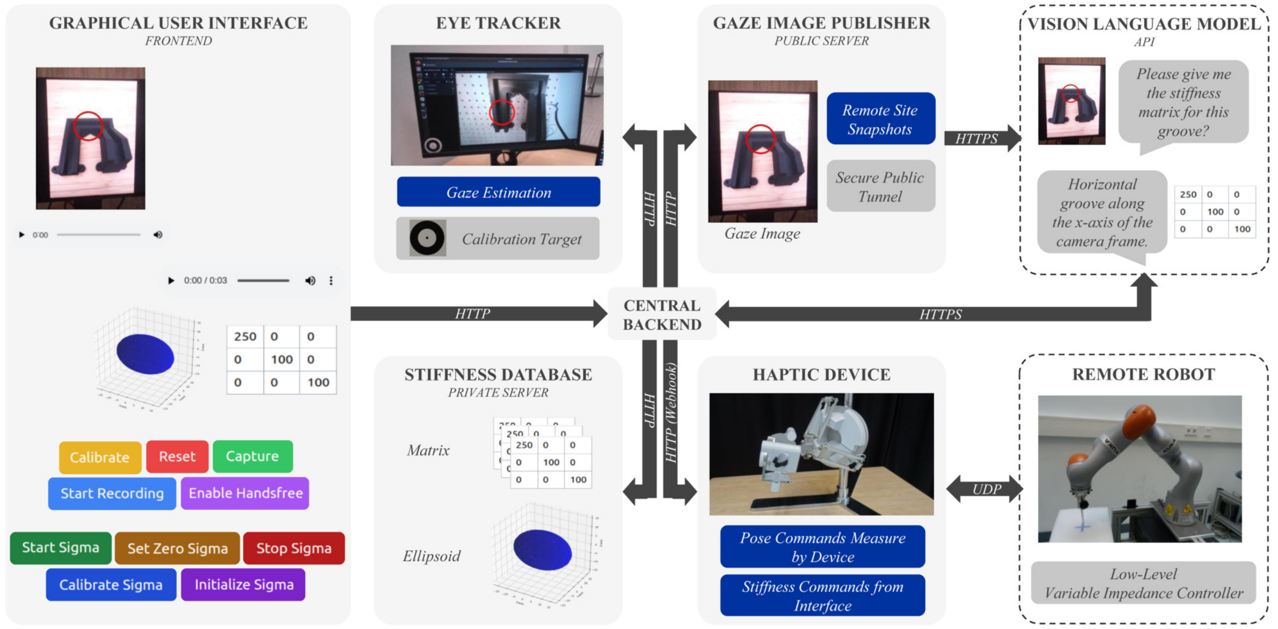


FIGURE 2
The diagram shows the container-based software architecture of the visio-verbal teleimpedance interface, emphasising the communication links between the different containers. The Arrows indicate the data flow between these containers, and the blue blocks indicate the information that is passed. The blocks surrounded by a dotted line are not containers.

This information is collected by the *Central Backend* to create a conversation history for the VLM. It also offers speech-to-text and text-to-speech capabilities, which are utilised in the process. The request is then sent to the VLM API. Since our VLM operates on external public servers, the *Gaze Image Publisher* container takes snapshots of the remote site overlaid with gaze estimates represented as red circles, and with an internet-accessible tunnel allows the online VLM to retrieve gaze images through web-accessible URLs. The stiffness received from the VLM is then sent to the GUI by the *Central Backend* to visually inform the operator of the current impedance configuration.

Finally, the stiffness is sent to the *haptic device*, where it is first combined with measured operator motion to create a unified command to be then sent to the remote robot. This container manages the haptic interface and ensures seamless communication with the remote robot controller via UDP communication, handling both pose commands and real-time stiffness updates. To improve safety, it includes a start/stop control mechanism that prevents unintended robot movements when the operator is not actively engaged. Additionally, using the GUI, the operator can automatically align the haptic device's zero position with the robot's current position, avoiding sudden jumps and ensuring smooth operation. This also enables workspace re-indexing during the operation.

3 Experiments

To validate the proposed interface and demonstrate the key functionalities, we conducted experiments on a real teleoperation setup (see Figure 3). At the local site, the operator monitors the remote environment via a display screen, which shows the live feed from a camera mounted at the end-effector of the remote Kuka LBR iiwa robotic arm. The operator commands the remote robot reference position through a Force Dimension Sigma.7 haptic device. The impedance adjustments are determined based on the operator's gaze, recorded by an eye-tracking system (Tobii Pro Glasses 2), and the operator's verbal interaction, captured by the built-in microphone on the laptop. The laptop also serves as the central processing unit for gaze and verbal input. The containers that run on it are created with *Docker*, and the secure public tunnel to make the images available through public URLs is done with *ngrok*. Finally, the VLM used is GPT-4o. To ensure transparency and confirm successful impedance adjustments, the system provides multimodal feedback. The selected stiffness matrix is conveyed through both a verbal notification via the laptop's speakers and a visual representation in the form of a stiffness ellipsoid displayed on the GUI.

For the experiment task, we selected a *slide-in-the-groove* task, which is common in assembly and involves physical interaction in multiple axes. We used a 3D-printed structure as seen in Figure 4, where the robot had to insert a peg into the groove and then navigate it while maintaining low interaction forces. The ability to dynamically adjust stiffness is critical to ensure smooth insertion, and subsequent traversing through the grooves without damage to the structure or the peg.

To perform this task optimally, the stiffness properties must be adapted across different phases of the task. At the entrance of the groove, the robot arm must be lowered down into the groove and align itself with the groove entrance without exerting excessive force on the structure. This is achieved by commanding a low stiffness

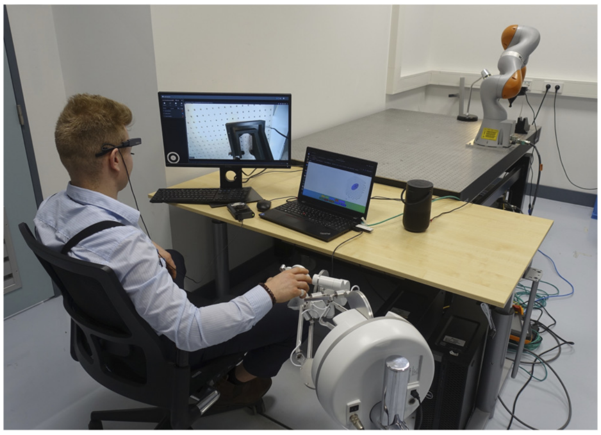


FIGURE 3
Experiment setup with teleimpedance system. The left screen provides a real-time camera feed mounted on the remote robot end-effector. The operator wears eye-tracking glasses and operates a haptic device to control the motion of the remote robot. The visio-verbal interface takes estimated gaze and verbal interaction as input and outputs the commanded stiffness for the remote robot based on the given task context.

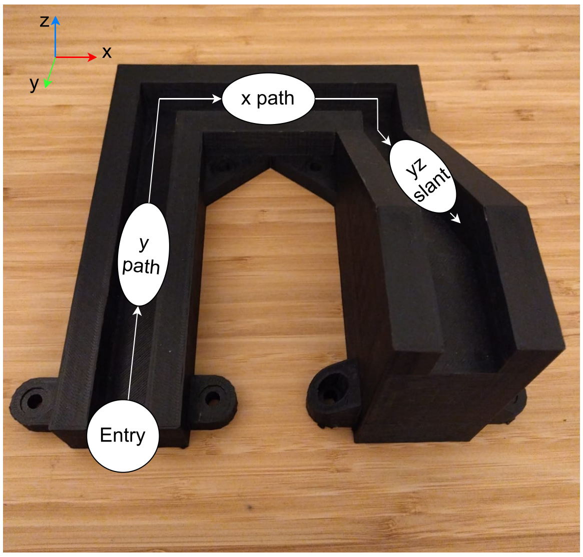


FIGURE 4
Groove structure with ellipsoid shapes (white) indicating the desired stiffness matrices for each phase of the task.

in the x-axis and y-axis, allowing compliant self-alignment in that plane, while commanding *high stiffness* in the z-axis to have the strength to push the peg inside. Once inside, the stiffness in the direction of the groove (y-axis) must be increased to ensure friction is compensated for accurate tracking of the reference trajectory. Meanwhile, stiffness in the directions of walls and bottom (x-axis and z-axis) should be low for the peg to comply with the environmental constraints and prevent excessive normal forces that could lead to jamming or structural damage. When the groove changes direction, the robot stiffness ellipsoid should be adjusted in a similar manner.

The experimental work was divided into several experiments. In the first experiment, we performed prompt optimisation for the interface to explore different parameter settings and find the best one to be used for subsequent demonstration experiments. The purpose of the second experiment was to show verbal aspects of the interface, focusing on demonstrating fluid task execution. The third experiment focused on demonstrating the gaze and conversational aspects.

3.1 Experiment 1: prompt parameter optimisation

The goal of the prompt parameter optimisation study, specific to the use of VLM in teleimpedance, was to investigate and compare the effects of different settings (system role, amount of priors, and image resolution) on the stiffness matrix predictions. VLMs extend the capabilities of large language models (LLMs) by incorporating image processing alongside textual input. A key discovery was their ability to perform few-shot learning, where instead of fine-tuning a large model for a specific task, a small set of example question-answer pairs alongside the user's prompt are provided (Brown et al., 2020). Studies have shown that few-shot prompting significantly improves VLM performance compared to zero-shot prompting, in which the model is asked to complete a task without prior examples (Alayrac et al., 2022).

Our research investigates an extension of few-shot learning in VLMs by incorporating gaze-estimate images alongside a curated set of labelled examples that define the corresponding stiffness matrices. This approach leverages the few-shot paradigm to enhance the model's ability to infer appropriate stiffness configurations based on visual and verbal contextual cues. The way the VLM is prompted significantly influences the quality of its outputs. To ensure optimal performance by generating accurate stiffness matrices in different phases of the task, we conduct a parameter optimisation experiment. This allowed us to identify the most effective combination of prompt parameters for teleimpedance application, including the task description, example demonstrations, and image detail.

In deep learning terminology, the *system role* directs the VLM to a specific function, which was, in our case, a specialised stiffness matrix generator. This ensured that operator input was interpreted in terms of matrix updates rather than just general conversation. We evaluated three prompt designs (system roles) with progressively stronger inductive bias, each building on the previous one. Role 1 provides only an instruction to generate a stiffness matrix from the most recent input image, the required response format, and permission to use conversation history (i.e., few-shot examples of image–stiffness matrix pairs). Role 2 introduces the task physics as an impedance-control problem (virtual spring aligned with the groove), adds a solution procedure, and specifies the camera frame and numeric guidance (e.g., compute stiffness for the highlighted groove; prescribe high stiffness 250 along the groove and low stiffness 100 in orthogonal directions and express it in terms of the Cartesian camera frame). Building on Role 2, Role 3 further includes textual labels for specific groove sections with their corresponding target stiffness matrices (e.g., “groove along X-axis (left–right)” corresponds to diag(250, 100, 100)).

Another important aspect is related to the few-shot demonstration approach and the inclusion of conversation *priors*. These contain example prompts (e.g., “What is the stiffness matrix for this phase?”) with corresponding images for different phases of the task paired with the desired outputs in the form of stiffness matrices for the given phases of the task. Since we had four phases of the task (entrance, y-traverse, x-traverse, and yz-slant), four example prompts and four corresponding images were fed to the VLM. This list was tested under three conditions. In the None condition, the model was tested without any example prompts and responses. In the Ideal condition, the prior message list was created from a setup in the environment, which had less challenging lighting and camera angle than the lab environment. In the Lab condition, the prior message list was created from the lab environment, which had challenging lighting and camera angle. The inclusion of the Ideal condition allowed us to investigate whether the model’s performance would improve when provided with prior data from either the ideal or lab environment, shedding light on how environmental differences influence the results.

Finally, in image pre-processing, the images were overlaid with the gaze estimate before providing them to the GPT-4o API. The image detail was varied between low and high settings to determine its impact on performance. The *low-detail* mode provided faster responses, while the *high-detail* mode allowed the model to process finer details in the input images. From experiment 1, high detail had an increased performance in stiffness matrix generation and was therefore implemented in experiment 2. In our system, the input images with gaze estimates had a resolution of 1920 × 1080 pixels. This image was then scaled to 768 × 1364 pixels by the API while keeping the aspect ratio the same. The API then determines how many tiles of 512 × 512 pixels it needs to fill the image 768 × 1364 pixels, which is six in this case. Therefore, in high resolution mode, the API receives a 512 × 512 pixels down-sample of the whole image with budgeted 85 tokens for image description and another six detailed patches of 512 × 512 pixels covering the whole image of 768 × 1364 pixels, where each patch is budgeted to 170 tokens for image description. In the low-detail mode, the API receives only a down-sampled image of 512 × 512 pixels with budgeted 85 tokens for image description.

The conversation history maintains a record of the operator’s previous commands and context, including gaze snapshots, enabling iterative refinement of stiffness matrices. For the tests in Experiment 1, we had only priors and no additional conversation history to make sure that the comparison between conditions was fair. In the subsequent demonstration experiments (i.e., 2 and 3), we let the VLM keep the conversation history to continue improving and learning. Since more history also makes VLM slower, we limited it to 10 prompts to ensure faster responses.

In summary, every prompt contains three elements: re-purposing, conversation primer (few-shot learning), and conversation history. For the re-purposing element, we set the system role definition for the API prompt of the VLM (GPT-4o) to one of the roles 1, 2 or 3. For the conversation primer (few-shot learning) element, we used images with gaze estimates overlaid on top of them, together with a request for the stiffness matrix for that task phase. We did this by filling in the “user” part of the conversation with this data. We then filled in the “system” or response part with the correct answer, which is embodied as the correct stiffness matrix for the part of the task highlighted with the gaze estimate in the image provided by the user. Just like the re-purposing element of the prompt, the conversation primer element is static and does not change over time. Therefore, the explicit prompt text contained the re-purposing element (role 1, 2 or 3) together with one version of the conversation primer (None, Ideal or Lab) as its first two blocks.

Finally, the conversation history resulted from the actual prompting of the GPT-4o API, where each prompt would start with the first two blocks and continue with a request by the user in the form of a question (e.g., “What is the stiffness matrix for this part?”) and an image overlaid with the gaze estimate. While the system role and the conversation primer are constant over time, the conversation history changes over time and develops just like any transcript of any human conversation over time would. Therefore, this element of the explicit prompt text grows over time as the user continues to supply images of the task with gaze estimates and questions. For the streamlined verbal mode of the interface (experiment 2), this element consisted only of a single prompt (image with gaze estimate and question). The code then extracted the stiffness matrix from the response, where the stiffness matrix conversion to the ellipsoid representation was based on eigen-decomposition. For the full mode combining gaze and conversational aspects (experiment 3), this element would be increasing in size over time while the operator progressed through the task, as each phase of the task would require the user to provide a new image with a gaze estimate and stiffness matrix request, and it would require the GPT-4o API to respond to that.

In the prompt optimisation experiment, we first performed a pre-test on the prediction success rate of all possible combinations with 5 trials for each combination to eliminate the worst-performing ones. In each trial, we tested whether the VLM could correctly generate the stiffness matrix for the four different phases of the task (entrance, y-traverse, x-traverse, and yz-slant). Figure 5 shows the prediction success rate results of the pre-test for the four different phases of the task for all possible combinations. The prediction success rate was calculated as the percentage of correct stiffness predictions out of the total trials. The correct stiffness was defined based on the optimal strategy for a given phase of the task, as previously defined above. For example, during sliding, the correct stiffness was defined to be high in the direction of pushing, and low stiffness in the direction to the walls.

For the experiment, we selected a challenging scene with difficult lighting and camera angle in the lab environment. Note that since the camera angle was directly above the structure, it was impossible for the vision to identify the upward slant, the results for the slat phase are accordingly poor. Nevertheless, for the other phases of the task where vision could detect the structure, VLM achieved an excellent success rate.

Following the pre-test, we selected the nine best-performing combinations for more extensive testing, where we used 15 trials for each combination. We kept the setting from the pre-test with a challenging scene with difficult lighting and camera angle. Table 2 shows the results of the main test with

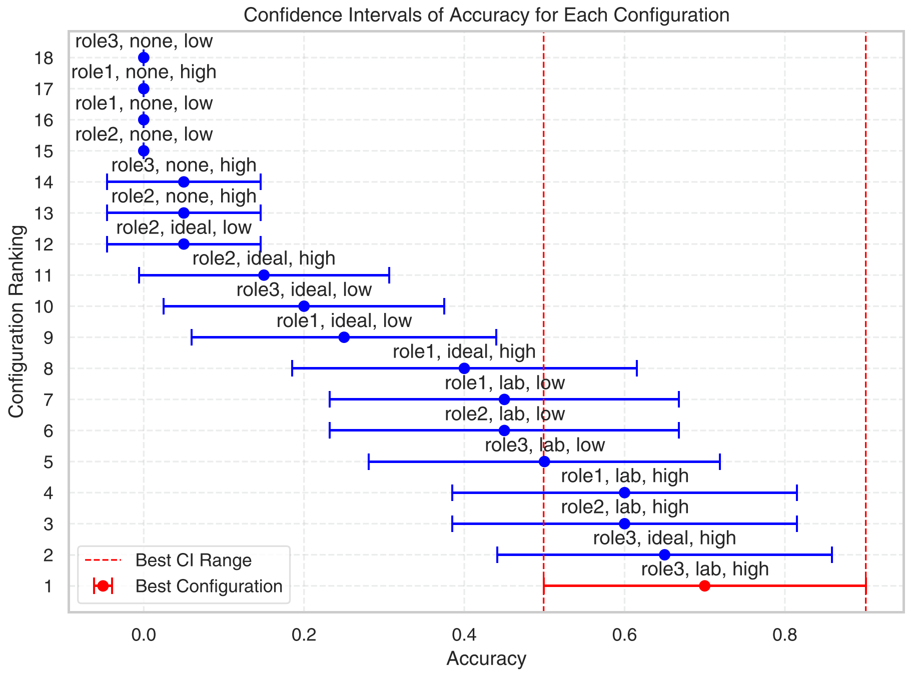


FIGURE 5
Results of a pre-test for all prompt configurations with means (dots) and confidence intervals (whiskers). Prediction accuracy was measured in terms of prediction success rate calculated over 5 trials for each combination. The best nine configurations from the pre-test were used for the main test on more trials in Table 2.

TABLE 2 Accuracy of stiffness matrix prediction for the main test expressed as prediction success rate (results above 90% are highlighted by bold text). The success rate for each individual phase is calculated over 15 trials, while the overall success rate is determined from trials of all sections combined (60 trials). The last column reports the overall success rate excluding the slant phase (45 trials). The first three columns indicate system role, type of prior, and camera image resolution, where each row represents a unique tested combination.

Role	Prior	Resolution	Entrance	Y-traverse	X-traverse	YZ slant	Overall (Slant)	Overall (No slant)
3	Lab	High	**1.00**	**0.93**	**1.00**	0.00	0.73	**0.98**
1	Lab	High	**1.00**	0.67	**0.93**	0.13	0.68	0.87
2	Lab	High	**0.93**	0.67	**1.00**	0.07	0.67	0.87
3	Ideal	High	**0.93**	0.67	**0.93**	0.00	0.63	0.84
1	Lab	Low	**1.00**	0.27	**1.00**	0.00	0.57	0.76
2	Lab	Low	**1.00**	0.00	**1.00**	0.00	0.50	0.67
3	Lab	Low	**1.00**	0.00	**1.00**	0.00	0.50	0.67
1	Ideal	High	0.73	0.40	0.60	0.00	0.43	0.58

the average prediction success rate based on entrance, y-traverse, x-traverse, and yz slant phases. Based on the insight from this experiment, we determined that the optimal prompt configuration was: an elaborate task description with labelled examples (Role 3) for the system role, the inclusion of lab-based exemplars as a primer (Lab), and high-detail image processing (High). This refined configuration was then used in the subsequent demonstration experiments.

3.2 Experiment 2: Demonstration of verbal commands

This experiment aimed to demonstrate the interface in a case when the operator wants a fluid task execution with minimum interaction with the interface. The operator issues purely verbal commands (e.g., "Increase stiffness along the groove axis") without providing gaze snapshots. This demonstration verifies that the

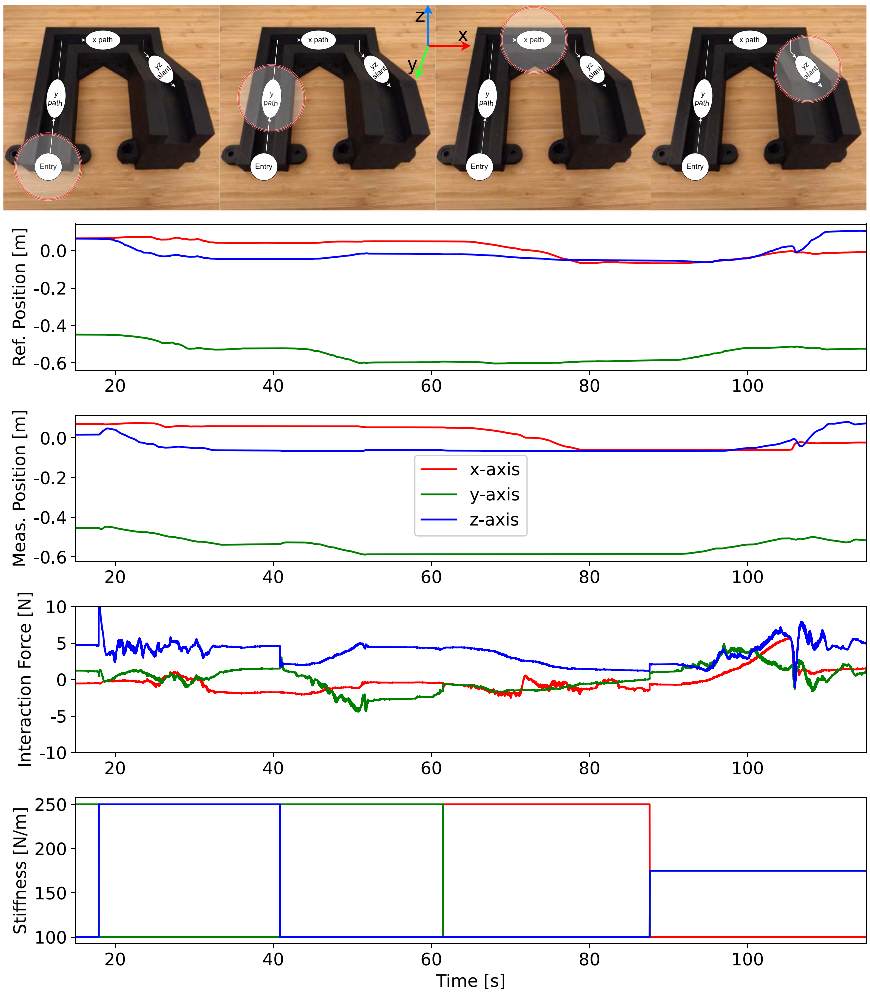


FIGURE 6
Results of the experiment with swift verbal commands. The images on top show key phases of the task in the groove structure, including entrance, y-traverse, x-traverse and yz slant. The graphs show reference (first) and measured (second) robot end-effector position, measured forces (third), and stiffness (fourth). Different colours indicate variables in different axes. The stiffness changes also indicate transitions between phases.

interface can still produce meaningful matrices when images are absent, although it relies on user-specified axis names rather than visual context.

The results are presented in Figure 6. We can see the different phases of the task from the series of images on top. The robot reference movement as commanded by the operator is seen in the first graph, and the corresponding actual movement of the robot is shown in the second graph. The fourth graph shows the robot stiffness adjustments by the interface based on the operator's verbal command.

When the operator informed the interface about the intention to enter the groove, the stiffness was adjusted in a way that the robot became stiff along the z-axis and compliant along the other axes. This corresponds to a standard peg-in-the-hole strategy, where the robot should be stiff in the direction of pushing and compliant elsewhere to allow the peg to comply with the environment and slide inside. By observing the blue line at around 20 s, we can see how the operator then performed the insertion of the peg into the groove along the vertical direction (z-axis).

In the next phase, the operator communicated to the interface the intention to traverse along the y-axis inside the groove. The stiffness was changed so that the robot became stiff along the y-axis and compliant along the other axes. This ensured that the peg overcame the friction needed to move in the y-axis, while interaction forces with the rigid wall were kept low. Green lines on the first and second graphs at around 45 s show the robot movement that the operator performed along the groove in the y-axis.

A similar process can be observed for movement in the groove along the x-axis around 65 s, where stiffness was changed to be high in the x-axis. In the final phase, the movement in the groove was slanted upward and when the operator informed the interface about the intention, the stiffness was adjusted to be stiff along 45° in the y-z plane. The final slanted movement of the robot can be seen around 100 s.

The third graph shows that the interaction force on average stayed below 5 N, with only one short peak around 10 N. This interaction force performance is comparable to a study in the literature, where a hand-held manual teleimpedance interface was used for the robot control to navigate the same structure (Kraakman and Peternel, 2024).

3.3 Experiment 3: demonstration gaze and conversational aspects

This experiment aimed to demonstrate the interface with combined gaze and conversational aspects. While the focus of the previous experiment was on fluid task execution with the operator providing swift verbal commands, here the focus is on rich visio-verbal interaction. The gaze estimate and scene snapshot are included, allowing the operator to state the intentions (e.g., "I want to enter the structure"), while the VLM uses both the gaze-labelled snapshot and voice transcription to generate a more contextualised matrix.

The results are presented in Figure 7. The different phases of the task are illustrated with a series of images on top. The graphs show the same variables as in the previous experiment. The conversational history between the operator and the interface in each phase of the task is also presented in the figure. In the first phase, the operator looked at the entrance and communicated the intention to enter the structure to the interface verbally. Through the detected gaze and verbal context, the interface generated the correct stiffness matrix to enter the groove and gave the operator a verbal confirmation with some extra details. After the operator entered the groove with the peg on the remote robot end-effector, the gaze was changed to the next phase and a new verbal context related to the movement along that section was given to the interface. After the verbal confirmation, the operator moved along the y-axis. A similar procedure was done for the next phase, which involved movement along the x-axis.

At that point, the operator wanted to show the advantage of conversation history in the interface by reversing the movement to backtrack along the structure. Instead of communicating specific intentions regarding each phase, the operator just instructed the interface that he wanted to backtrack. The interface then sequentially changed the robot stiffness matrix when backtracking the previous phases until the operator exited the structure at the point it was entered. In case the interface made a mistake in classification and generated an incorrect stiffness matrix, the operator could use verbal communication to provide corrections.

4 Discussion

The results show that the visio-verbal teleimpedance interface can generate adaptive stiffness matrices under dynamic verbal interaction and visual cues. Experiment 1 provided insights into prompt engineering that optimised for teleimpedance applications. Experiment 2 demonstrated the use case where the operator can use the proposed interface for fluid task execution with minimal visio-verbal interaction, at the expense of advanced functionalities. Experiment 3 showed the full potential of visio-verbal interaction with rich visio-verbal interaction, where the operator could naturally indicate things by gaze and also receive verbal confirmations.

The experiments also highlighted the limitation of using a single camera angle as an input into VLM. For complex structures like the one used in this study, some features may be difficult to extract by vision. For example, in a top-down camera view and poor lighting conditions, the slant in the structure can be impossible to recognise. In this camera angle, it was simply misclassified as y-traverse. We performed a post-experiment test with a slightly easier camera angle and better lighting conditions (i.e., the one seen in Figure 4). In that case, the prediction success rate for slant increases to 66.6%.

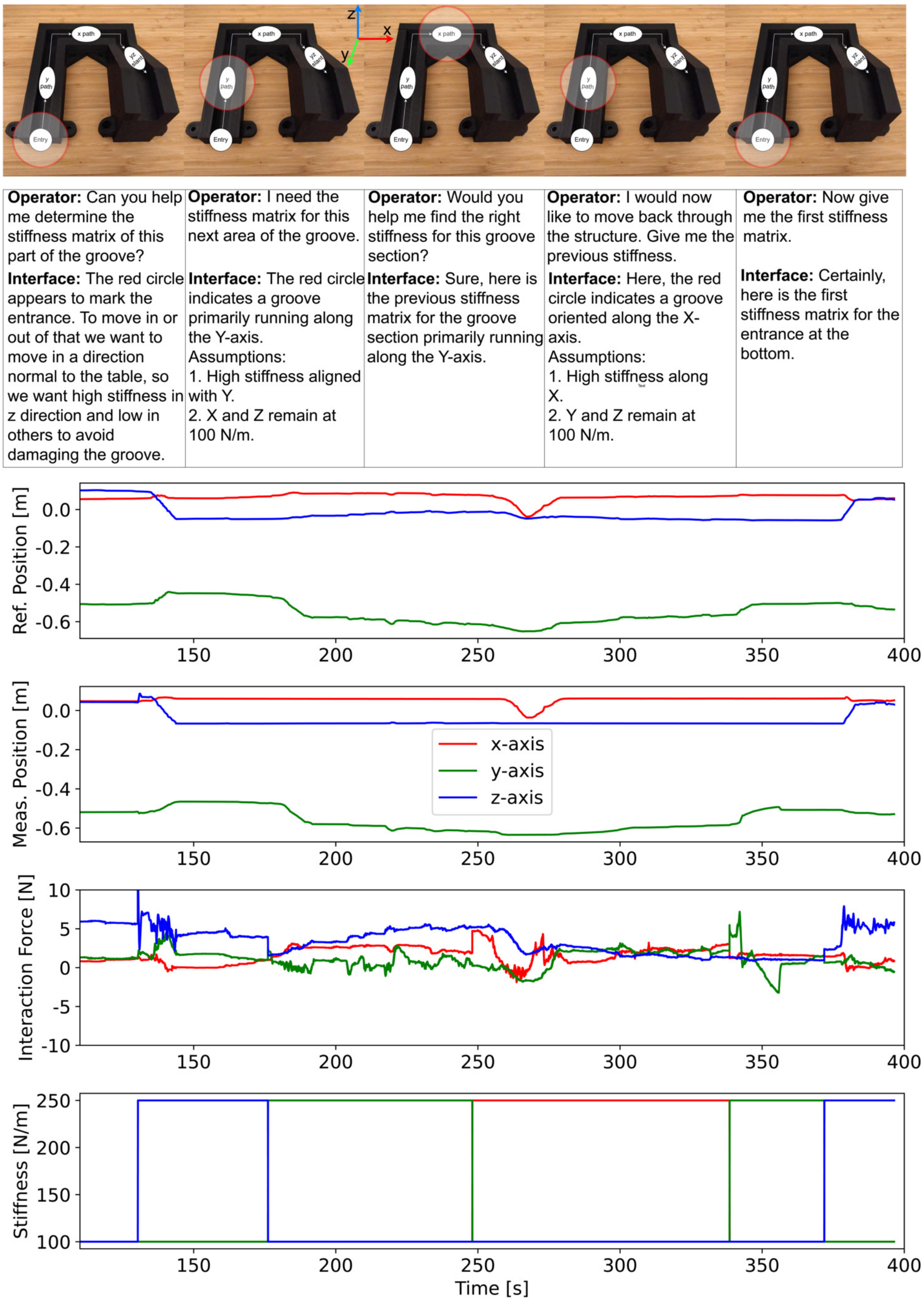
To resolve the camera angle issue and improve generality, one solution could be to install multiple cameras in the workspace to enable multi-view imagery, which would come at the expense of setup complexity. Moreover, in inaccessible remote environments, it would not be possible to set up fixed cameras around the scene, and a robot-mounted camera would be the only option. In that direction, we could design an autonomous active camera control method, where the robot would take images of the scene from multiple angles. This could also enable to make explicit 3D reconstruction of the environment. However, this might interrupt the operation, requiring some time to scan the scene. Another solution could be to use sensor fusion, such as exploiting depth information or even combining visual information with haptic information. However, haptic sensing requires a reactive approach, while the strength of setting impedance in advance of physical interaction is the ability to be proactive.

The prompt optimisation experiment provided some additional insights regarding the teleimpedance application. For example, the image resolution has a major effect on the prediction success rate. As seen in Table 2, four top-performing configurations all have high-resolution images. Another interesting insight is that including image priors with very good lighting/camera conditions that are from a different environment (Ideal) than the one used in the actual task (Lab) does not help the performance of the teleimpedance interface. As seen in Table 2, only two configurations in the nine top-performing had Ideal priors, and the top three all had Lab priors.

In the experiments, the stiffness matrices were diagonal since the principal axes of the task frame (i.e., groove structure) and the robot base frame were aligned. If they were not aligned, the visio-verbal interface could still generate the stiffness matrices in the task space. Then, methods such as eigen-decomposition or singular value decomposition could be employed to transform them into the base frame, where the robot impedance control is operating. This approach is much simpler and more effective than having VLM operate in the robot base frame and directly generate non-diagonal matrices.

One potential limitation of this study is the reliance on a mobile eye tracker with glasses, which necessitates additional image-processing steps to isolate the display screen from the surrounding environment. An alternative would be to employ a static eye tracker on the display screen itself.

The scope of this study was the design of the visio-verbal teleimpedance interface, analysis of the prompt optimisation,



Operator: Can you help me determine the stiffness matrix of this part of the groove?	**Operator:** I need the stiffness matrix for this next area of the groove.	**Operator:** Would you help me find the right stiffness for this groove section?	**Operator:** I would now like to move back through the structure. Give me the previous stiffness.	**Operator:** Now give me the first stiffness matrix.
Interface: The red circle appears to mark the entrance. To move in or out of that we want to move in a direction normal to the table, so we want high stiffness in z direction and low in others to avoid damaging the groove.	**Interface:** The red circle indicates a groove primarily running along the Y-axis. Assumptions: 1. High stiffness aligned with Y. 2. X and Z remain at 100 N/m.	**Interface:** Sure, here is the previous stiffness matrix for the groove section primarily running along the Y-axis.	**Interface:** Here, the red circle indicates a groove oriented along the X-axis. Assumptions: 1. High stiffness along X. 2. Y and Z remain at 100 N/m.	**Interface:** Certainly, here is the first stiffness matrix for the entrance at the bottom.

FIGURE 7
Results of the experiment with combined gaze and conversational aspects. The layout is the same as Figure 6. Additionally, it lists the conversation history between the operator and the interface in each phase of the task. The stiffness changes also indicate transitions between phases.

and technical evaluation of key functionalities on a real teleoperation setup. The next step is to design and conduct a human factor experiment to gather user experience insights and make a comparison analysis to alternative teleimpedance interfaces from the state-of-the-art.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

HJ: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review and editing. AD: Conceptualization, Formal Analysis, Investigation, Methodology, Supervision, Visualization, Writing – original draft, Writing – review and editing. LP: Conceptualization, Formal Analysis, Investigation, Methodology, Resources, Supervision, Visualization, Writing – original draft, Writing – review and editing.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Acknowledgements

The authors would like to thank Renchi Zhang for the help and advice regarding the eye tracking system and Yke Bauke Eisma for lending the eye tracking system.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2026.1749105/full#supplementary-material

References

Admoni, H., and Srinivasa, S. S. (2016). "Predicting user intent through eye gaze for shared autonomy," in *AAAI fall symposia*, 298–303.

Ajoudani, A., Tsagarakis, N., and Bicchi, A. (2012). Tele-impedance: teleoperation with impedance regulation using a body–machine interface. *Int. J. Robotics Res.* 31, 1642–1656. doi:10.1177/0278364912464668

Ajoudani, A., Fang, C., Tsagarakis, N., and Bicchi, A. (2018). Reduced-complexity representation of the human arm active endpoint stiffness for supervisory control of remote manipulation. *Int. J. Robotics Res.* 37, 155–167. doi:10.1177/0278364917744035

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., et al. (2022). Flamingo: a visual language model for few-shot learning. *Adv. Neural Information Processing Systems* 35, 23716–23736.

Aronson, R. M., Almutlak, N., and Admoni, H. (2021). "Inferring goals with gaze during teleoperated manipulation," in *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (IEEE), 7307–7314.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. *arXiv Preprint arXiv:2005.14165*. doi:10.48550/arXiv.2005.14165

Buscaglione, S., Tagliamonte, N. L., Ticchiarelli, G., Di Pino, G., Formica, D., and Noccaro, A. (2022). "Tele-impedance control approach using wearable sensors," in *2022 44th annual international conference of the IEEE engineering in medicine & biology Society (EMBC)* (IEEE), 2870–2873.

Cui, Y., Zhang, Y., Tao, L., Li, Y., Yi, X., and Li, Z. (2025). End-to-end dexterous arm-hand vla policies via shared autonomy: vr teleoperation augmented by autonomous hand vla policy for efficient data collection.

Díaz Rosales, A., Rodriguez-Nogueira, J., Matheson, E., Abbink, D. A., and Peternel, L. (2024). *Interactive multi-stiffness mixed reality interface: controlling and visualizing robot and environment stiffness. In 2024.* IEEE/RSJ International Conference on Intelligent Robots and Systems IROS, 13479–13486. doi:10.1109/IROS58592.2024.10801866

Doornebosch, L. M., Abbink, D. A., and Peternel, L. (2021). Analysis of coupling effect in human-commanded stiffness during bilateral tele-impedance. *IEEE Trans. Robotics* 37, 1282–1297. doi:10.1109/TRO.2020.3047064

Fu, Z., Song, P., Hu, Y., and Detry, R. (2025). Tasc: task-aware shared control for teleoperated manipulation. doi:10.48550/arXiv.2509.10416

Garate, V. R., Gholami, S., and Ajoudani, A. (2021). A scalable framework for multi-robot tele-impedance control. *IEEE Trans. Robotics* 37, 2052–2066. doi:10.1109/tro.2021.3071530

Gourmelen, G., Navarro, B., Cherubini, A., and Ganesh, G. (2021). *Human guided trajectory and impedance adaptation for tele-operated physical assistance. In 2021.* IEEE/RSJ International Conference on Intelligent Robots and Systems IROS, 9276–9282.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., et al. (2020). Scaling laws for neural language models. *arXiv Preprint arXiv:2001.08361*. doi:10.48550/arXiv.2001.08361

Konrad, R., Padmanaban, N., Buckmaster, J. G., Boyle, K. C., and Wetzstein, G. (2024). GazeGPT: augmenting human capabilities using gaze-contingent contextual AI for smart eyewear. *arXiv Preprint arXiv:2401.17217*. doi:10.48550/arXiv.2401.17217

Kraakman, M. F., and Peternel, L. (2024). "Design and evaluation of finger-operated teleimpedance interface enabling simultaneous control of 3d aspects of stiffness ellipsoid," in *2024 IEEE-RAS 23rd international conference on humanoid robots (Humanoids)* (IEEE), 690–697.

Liu, H., Shah, R., Liu, S., Pittenger, J., Seo, M., Cui, Y., et al. (2025). Casper: inferring diverse intents for assistive teleoperation with vision language models. *arXiv Preprint arXiv:2506*. doi:10.48550/arXiv.2506.14727

Michel, Y., Rahal, R., Pacchierotti, C., Giordano, P. R., and Lee, D. (2021). Bilateral teleoperation with adaptive impedance control for contact tasks. *IEEE Robotics Automation Lett.* 6, 5429–5436. doi:10.1109/lra.2021.3066974

Naceri, A., Schumacher, T., Li, Q., Calinon, S., and Ritter, H. (2021). Learning optimal impedance control during complex 3d arm movements. *IEEE Robotics Automation Lett.* 6, 1248–1255. doi:10.1109/lra.2021.3056371

Park, S., Lee, W., Chung, W. K., and Kim, K. (2018). Programming by demonstration using the teleimpedance control scheme: verification by an semg-controlled ball-trapping robot. *IEEE Trans. Industrial Inf.* 15, 998–1006. doi:10.1109/tii.2018.2876676

Peternel, L., and Ajoudani, A. (2022). After a decade of teleimpedance: a survey. *IEEE Trans. Human-Machine Syst.* 53, 401–416. doi:10.1109/thms.2022.3231703

Peternel, L., Beckers, N., and Abbink, D. A. (2021). "Independently commanding size, shape and orientation of robot endpoint stiffness in tele-impedance by virtual ellipsoid interface," in *2021 20th international conference on advanced robotics* (Ljubljana, Slovenia: IEEE), 99–106.

Si, W., Wang, N., and Yang, C. (2021). A review on manipulation skill acquisition through teleoperation-based learning from demonstration. *Cognitive Comput. Syst.* 3, 1–16. doi:10.1049/ccs2.12005

Siegemund, G., Díaz Rosales, A., Glodde, A., Dietrich, F., and Peternel, L. (2024). "Semi-autonomous teleimpedance based on visual detection of object geometry and material and its relation to environment," in *2024 IEEE-RAS 23rd international conference on humanoid robots (Humanoids)*, 779–786. doi:10.1109/Humanoids58906.2024.10769858

Suomalainen, M., Karayiannidis, Y., and Kyrki, V. (2022). A survey of robot manipulation in contact. *Robotics Aut. Syst.* 156, 104224. doi:10.1016/j.robot.2022.104224

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30, 5998–6008.

Walker, D. S., Salisbury, J. K., and Niemeyer, G. (2011). "Demonstrating the benefits of variable impedance to telerobotic task execution," in *2011 IEEE international conference on robotics and automation (ICRA)*, 1348–1353.

Yang, C., Zeng, C., Fang, C., He, W., and Li, Z. (2018). A dmps-based framework for robot learning and generalization of humanlike variable impedance skills. *IEEE/ASME Trans. Mechatronics* 23, 1193–1203. doi:10.1109/tmech.2018.2817589