Visio-Verbal Teleimpedance:

A Gaze and Speech-Driven VLM Interface for Human-Centric Semi-Autonomous Robot Stiffness Control

H.A. (Henk) Jekel[1]

Supervisors
Dr. L. (Luka) Peternel [1]
A. (Alejandro) Díaz Rosales[1,2]

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

Visio-Verbal Teleimpedance:

A Gaze and Speech-Driven VLM Interface for Human-Centric Semi-Autonomous Robot Stiffness Control

by

H.A. (Henk) Jekel[1]

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Monday March 24th, 2025 at 11:00 AM.

Student number:	5609593	
Project duration:	November 10th, 2023 – March 24th, 2025	
Thesis committee:	Dr. L. (Luka) Peternel,	TU Delft, supervisor
	A. (Alejandro) Díaz Rosales	TU Delft/CERN, supervisor
	Dr. C. (Chris) Pek,	TU Delft

TU Delft

Preface

It has been quite a journey, and I would like to express my deepest gratitude to Luka Peternel for his invaluable guidance. His countless pieces of advice have been instrumental in shaping this work. If I had to recommend a thesis supervisor to anyone, I would choose Luka without hesitation. When I hear others, especially PhD students, share their struggles with supervision, I am reminded of how fortunate I was to have Luka. Always polite, he gave me the freedom to explore while providing insightful feedback that helped me recognize and learn from my mistakes in the most constructive way possible.

I would also like to thank Alejandro Díaz Rosales for his support. Despite being in Switzerland, he was always present in our meetings, offering valuable feedback that enriched my work. I am looking forward to finally meeting him in person during my defense.

Additionally, I would like to thank Chris Pek for serving as a committee member and for the incredible opportunity he gave me in the multi-disciplinary project, where I had the chance to be a teaching assistant. It was truly rewarding to support students in achieving their goals by providing hardware and software assistance for the MIRTE Master. I am also grateful to Martin Klomp for allowing me to contribute to the design and the manufacturing of the MIRTE Master, an experience that has been invaluable in expanding my knowledge and skills.

Lastly, I extend my heartfelt thanks to my family and friends for their unwavering support throughout this journey. They witnessed the entire process, from my excitement when I first got the eye tracker to work and showed them an image with a red circle, to the moment they could access a website from home and make real-time snapshots with my glasses, seeing inside my house while I wore them, all the way to the final demo videos. Their encouragement made this journey all the more meaningful.

H.A. (Henk) Jekel[1]
Delft, March 2025

Contents

1

Paper

Visio-Verbal Teleimpedance: A Gaze and Speech-Driven VLM Interface for Human-Centric Semi-Autonomous Robot Stiffness Control

Henk H.A. Jekel*
Supervised by Luka Peternel and Alejandro Diaz Rosales

***Abstract*—Three-year-old toddlers can effortlessly guide a toy train along a wooden track, whereas this slide-in-the-groove position tracking task requires a skilled operator using a teleoperated robot arm due to the lack of direct contact and force feedback. Although an autonomous robot can perform this task in a fixed setup, telerobotics is crucial for unknown environments where human control is essential, as humans provide the adaptability needed to handle unpredictable conditions. The introduction of torque-controlled motors and haptic devices has enhanced teleoperation by improving telepresence and immersion. Operators can perceive interaction forces through the primary position control input via the haptic device, while a secondary control input allows them to adjust the robot arm's impedance. This ability, known as teleimpedance, allows operators to control the robot's physical interaction based on environmental context. A toddler naturally remains relaxed in the plane perpendicular to the train's forward direction, where gravity and the groove sides provide stability, preventing derailment and wheel damage. At the same time, they maintain firmness a long t he t rack f or s mooth forward movement. Tele-impedance enables the operator to achieve a similar balance. It allows adaptation of an optimal balance of low and high impedance in different axes of Cartesian space to match task demands. Current impedance control interfaces rely on complex muscle activity measurements, requiring long calibration procedures to map the operator's arm stiffness to the robot arm. Other interfaces use hand-controlled input devices that must be operated in addition to the haptic device, reducing the operator's cognitive bandwidth for the position tracking task. Existing interfaces typically provide only partial stiffness control or introduce visual distractions. In contrast, we propose a novel visio-verbal interface that leverages gaze and speech, natural modes of interaction, to enable hands-free semi-autonomous control of translational stiffness in all three dimensions while maintaining visual attention on the position tracking task. The interface's vision-language model (VLM) determines the three-dimensional robot endpoint stiffness by combining the operator's verbal intent with gaze estimates from a mobile eye tracker. We demonstrate a proof of concept for this approach. The hardware includes Tobii Pro Glasses 2 eye trackers, a Force Dimension sigma7 haptic position input interface, and a KUKA LBR iiwa collaborative robot arm equipped with a custom-built endpoint camera mount for the Realsense D455 camera and a 3D-printed peg to evaluate the interface in a 3D-printed U-shaped slot for a slide-in-the-groove task similar to guiding a toy train.**

***Index Terms*—Teleimpedance, teleoperation, impedance control, gaze estimation, speech interaction, semi-autonomous control, vision-language model (VLM), human-centric robotics**

Delft University of Technology, Faculty of Mechanical Engineering, Department of Cognitive Robotics, Mekelweg 5, 2628 CD, Delft, The Netherlands

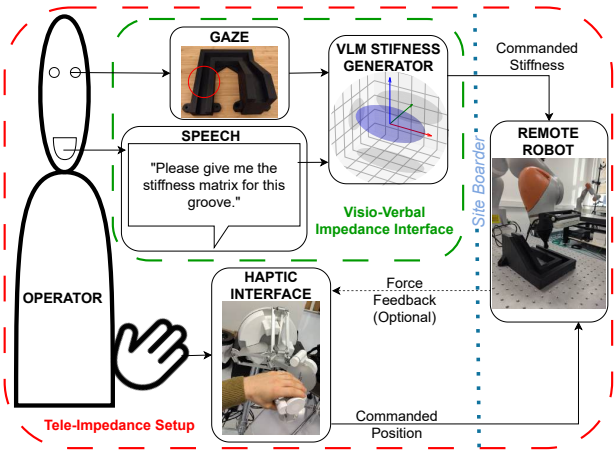

Fig. 1. Diagram illustrating the visio-verbal teleimpedance interface (green) within a teleimpedance setup (red). The operator controls the remote robot's position via a haptic device with optional force feedback (bottom) and adjusts impedance using gaze and speech input (top). The VLM processes these inputs to generate the appropriate stiffness matrix, which is directly applied to the robot. This introduces a novel impedance interface, enabling human-centric, semi-autonomous stiffness control.

I. Introduction

Telerobotics plays a crucial role in scenarios where remote human control is essential for safety and adaptability, such as disaster response [1], robot-assisted surgery [2] [3] [4], remote site inspection and maintenance [5],space [6] [7] and ocean [8] exploration and hazardous environment operations. [9] Although autonomous robots excel in structured environments found in manufacturing tasks, they struggle to adapt to dynamic and unstructured conditions due to limited cognitive flexibility [10]. To address this gap, teleoperation integrates human adaptability into robotic control, enabling operators at the local site to command robot movements at the remote site using interfaces such as haptic devices, joysticks, and motion capture systems. [10] This is similar to how a video game character mimics a player's actions, whether through a controller or motion tracking.

The integration of torque-controlled motors [11] [12] and haptic devices has significantly advanced teleoperation by improving telepresence and immersion through force feedback and impedance control. [13] To illustrate this, we refer to the general structure of the tele-impedance setup in Figure

1. From bottom left to right, the operator's hand movements are sensed by the haptic device and converted into position commands, which are sent to the robot arm's endpoint at the remote site, crossing the blue dotted site border. This one-way communication, where the operator sends position commands to the remote robot without receiving force feedback, is called unilateral teleoperation. With torque sensors, the haptic device can provide optional force feedback from the remote robot, as shown by the dotted arrow in Figure 1. This additional force feedback characterizes bilateral teleoperation, allowing the operator to perceive interaction forces. [14] At the top of the figure, teleoperation extends to teleimpedance [15] [16], where an additional control input from an impedance interface allows the operator to modulate the robot arm's impedance. This approach enables operators to adjust the robot's physical interaction based on the environment, improving performance in tasks that require variable stiffness or directionally dependent stiffness. For example, this additional input allows the operator to achieve a soft interaction when handling fragile materials like glass or to maintain stiffness only in the insertion direction for peg-in-hole assembly, facilitating self-alignment. [15] A detailed mathematical foundation of teleimpedance is given in appendix B

Existing teleimpedance command interfaces can be broadly categorized into manual impedance control and automated impedance control, depending on whether the human or the system determines the impedance of the interaction. [10] Manual control approaches allow the human operator to explicitly adjust stiffness through physiological signals [17] [18] [19], physical inputs [20] [21] [22], or touchscreen-based graphical interfaces. [23] One common method involves estimating the operator's stiffness using electromyography (EMG) sensors, which map muscle activity to the robot's impedance. [17]–[19] While EMG-based teleimpedance interfaces enable multi-degree-of-freedom stiffness control by estimating the endpoint stiffness ellipsoid of the human arm [17]–[19], they present several challenges. These systems require surface electromyography (sEMG) sensors to measure muscle activity, relying on antagonistic muscle co-activation to infer stiffness changes. [10] However, they demand precise calibration for a specific arm configuration, limiting their generalization across different poses. [10] Additionally, equipping multiple EMG electrodes is a tedious process, and even simplified approaches using fewer sensors or commercial arm braces still suffer from signal noise, motion artifacts, and cross-talk. [10] Moreover, while these methods effectively scale the size of the stiffness ellipsoid, studies suggest that they provide limited control over its shape. [24]

EMG interfaces are also susceptible to the coupling effect, where unexpected haptic feedback triggers involuntary reflexes in the operator's arm, leading to increased muscle activity. This unintended stiffening is detected by the EMG sensors and subsequently alters the commanded robot stiffness. Force grip sensor-based interfaces aim to reduce hardware complexity but still suffer from the coupling effect, as grip force is inherently linked to neuromuscular impedance [20]. Additionally, these interfaces can only control one degree of freedom per sensor, significantly restricting the operator's ability to shape the stiffness ellipsoid.

Alternative approaches introduce decoupled interfaces, where unintended stiffening due to force feedback does not influence the robot's endpoint stiffness. Examples include push-button interfaces [21] and tablet-based interfaces [23]. Push-button interfaces can only adjust a single degree of freedom at a time, limiting operator flexibility. Tablet-based interfaces offer a graphical representation of two planes of the stiffness ellipsoid, allowing adjustments through familiar touchscreen interactions. However, using a tablet while simultaneously performing a position-tracking task shifts the operator's focus away from the primary task, potentially disrupting workflow. A more recent approach to improving control over multi-degree-of-freedom stiffness is a single-handed 3D stiffness command device [22], which integrates two scroll wheels, a joystick, and a force sensor to enable more versatile stiffness control. While this device combines multi-degree-of-freedom impedance control with the ability to maintain visual attention on the primary task, it requires significant training and imposes a high cognitive load on the operator.

In contrast, automated impedance control systems remove the operator from the impedance control loop, allowing robot autonomy to determine the appropriate stiffness. For example, in [25], the robot used torque sensors to measure physical interaction forces and autonomously adjust stiffness for task stabilization, ensuring compliance when needed and increasing rigidity during disturbances. Similarly, [26] proposed a vision-based system where the robot detected objects and inferred their material properties to preemptively set optimal stiffness values. For instance, when approaching a fragile glass object, the system automatically reduced stiffness to avoid excessive contact forces. These fully autonomous approaches improve safety and reduce operator workload but also remove direct human input, reducing the operator's ability to intervene in uncertain environments. Both manual impedance control and automated impedance control are forms of the traded control paradigm where either the operator or the system is in full control of the robot impedance.

Despite advancements in teleimpedance interfaces, current methods either demand continuous manual adjustments, increasing cognitive load, or rely entirely on automation, reducing adaptability in uncertain environments. To bridge this gap, this work introduces a novel visio-verbal teleimpedance interface that follows the shared impedance control paradigm. As illustrated in Figure 1, focusing now on the proposed interface (green), it enables operators to adjust the robot's full 3D stiffness matrix, including its shape, size, and orientation, using gaze and speech, two natural and intuitive communication modalities. This approach eliminates the need for direct manual input while ensuring on-demand adaptability without diverting the operator's visual attention from the task and without the neuromechanical coupling effect.

Unlike previous methods that strictly required manual adjustments or complete automation, this system balances cognitive workload between the operator and a vision-language model (VLM). Figure 2 highlights how the system implements human-in-the-loop impedance control: the VLM generates a stiffness matrix based on the operator's verbal command

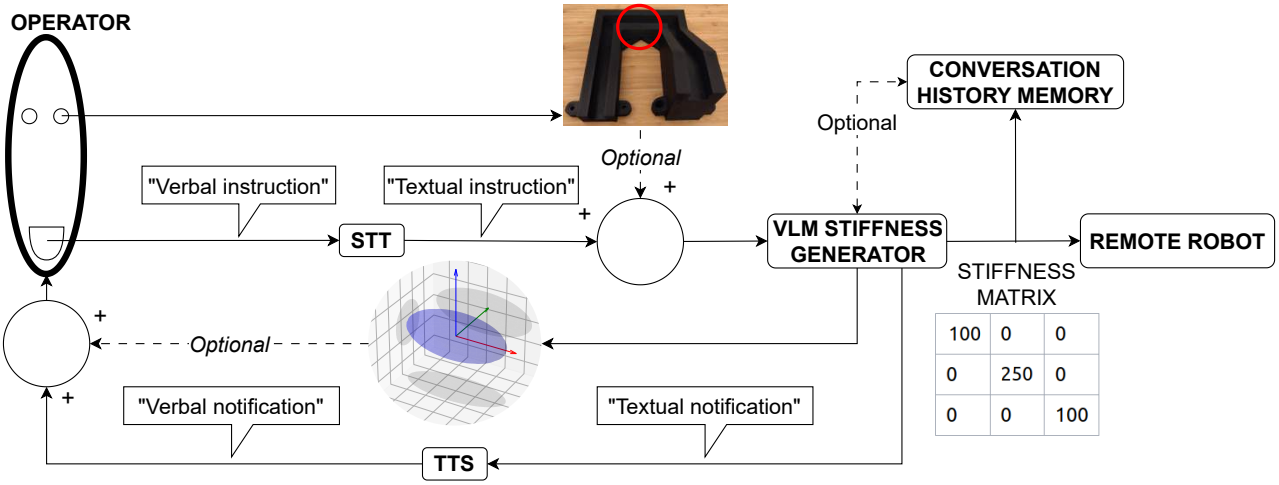

Fig. 2. Human-in-the-loop impedance control using verbal commands and scene images with overlaid gaze estimates. At the top, a snapshot of the teleoperation scene with gaze overlay is captured in parallel with the operator's verbal command (left), which is processed through a speech-to-text (STT) module. These inputs, along with relevant conversation history stored in memory, are provided to the vision-language model (VLM) for stiffness matrix generation. The generated stiffness matrix is then applied to the remote robot. Immediate feedback is given through text-to-speech confirmation and a real-time visualization of the stiffness ellipsoid, ensuring alignment with the operator's intent. If the operator disagrees with the output, they can modify it through subsequent verbal adjustments, optionally referencing specific entries in the conversation history to refine the VLM's response.

and optional gaze input, then communicates it via a brief verbal notification and an optional visual check via a stiffness ellipsoid (figure 8). If the operator disagrees with the proposed matrix, they can provide verbal instructions for refinement. Additionally, as shown in the top right of Figure 2, the system retains a conversation history, allowing operators to revert to previous stiffness configurations through simple voice commands. For example, this feature can assist when repeating a task in reverse or returning to a previous configuration after an adjustment. This human-in-the-loop design ensures stiffness adaptability in unstructured environments by preserving human intent, making it the first teleimpedance interface to integrate eye tracking and vision language models for shared impedance control.

Section II begins by outlining the design requirements, followed by a detailed presentation of the hardware and software architecture of the *visio-verbal* teleimpedance interface. It then describes the interaction with the VLM, focusing on the prompt design used in the experiments of Section III. Section III presents these experiments along with their results. Section IV evaluates this work in the context of existing teleimpedance research, discussing its contributions, strengths, and limitations. Finally, Section V summarizes the key findings, highlights the broader impact of this research, and suggests directions for future work.

II. Methods

For this work, a novel *visio-verbal* teleimpedance command interface was designed to enable hands-free adjustment of the orientation, size and shape of the 3D translational stiffness (8). This interface leverages eye-tracking and speech recognition to allow operators to specify stiffness parameters in a natural and intuitive manner without diverting their attention from the teleoperation task. By feeding these inputs into a vision-language model (VLM), the system ensures on-demand, adaptive stiffness control.

The structure and functionality of the impedance interface, highlighted in green in Figure 1, are further detailed in Figure 2, which illustrates the system's workflow from multi-modal input to stiffness adjustment. Starting from the top, it captures a snapshot of the teleoperation scene and overlays the operator's gaze estimate as a red circle. Simultaneously, starting from the left, the system processes the operator's verbal command through a speech-to-text (STT) module. The VLM then interprets these multi-modal inputs to generate a stiffness matrix, which is applied directly onto the remote robot. To facilitate verification, the system provides immediate feedback through both verbal confirmation using a text-to-speech model (TTS) and a real-time visualization of the updated stiffness in the form of a stiffness ellipsoid (figure 8), ensuring that adjustments align with the operator's intent.

This section is organized as follows. Section II-A defines the interface design requirements. Based on these requirements, Section II-B presents the design of the novel impedance command interface. Section II-C details the hardware setup, while Section II-D covers the software architecture. Finally, Section II-E explains the interaction with the VLM through prompting.

A. Interface Design Requirements

Several teleimpedance command interfaces were briefly discussed in the introduction, along with their limitations. Based on these insights and a critical analysis of related work [10], a set of design requirements was formulated to guide the development of the novel *visio-verbal* teleimpedance command interface. These requirements aim to address existing challenges and prioritize a human-centric design, focusing

on adapting the system to the user rather than forcing the user to conform to the system. The requirements are as follows:

- R1: Combine eye-tracking with verbal interaction for stiffness matrix generation.
- R2: Enable the operator to control the full 3D translational endpoint stiffness.
- R3: Allow stiffness adjustments without diverting visual attention from position tracking.
- R4: Minimize setup and calibration procedures.
- R5: Prevent coupling between force feedback and commanded stiffness.

These requirements address the challenges of existing teleimpedance interfaces and ensure a balance between intuitive control and adaptability.

Existing impedance control methods fall into two categories: manual and fully automated approaches. While manual methods provide flexibility, they require continuous operator input, increasing cognitive load. On the other hand, fully automated approaches optimize performance but remove human adaptability, making them unsuitable for uncertain environments. R1 ensures that the interface follows a shared control paradigm, leveraging a vision-language model (VLM) to process gaze and verbal inputs. This allows the operator to remain actively involved in impedance adjustments. Unlike manual systems that require direct parameter tuning [17]–[19], or fully autonomous systems that remove the human from the loop [25], [26], the proposed approach balances automation with human intent.

For effective teleimpedance control, the operator must have full access to the 3D stiffness matrix. Many existing interfaces only allow partial control. Tablet-based methods restrict adjustments to 2D planes [23], while EMG-based systems scale stiffness magnitude but offer limited ellipsoid orientation control [24]. R2 ensures that operators can freely adjust stiffness along all three translational degrees of freedom, providing intuitive and complete control over impedance.

Another limitation of traditional teleimpedance interfaces is the need for operators to shift their focus between control inputs and the teleoperation task. Tablet interfaces and other visual feedback methods force the operator to look away from the robot [23], disrupting workflow and increasing error potential. R3 addresses this by integrating gaze and speech, allowing seamless stiffness adjustments without breaking visual attention. The interface uses quick verbal and visual confirmations to ensure the operator stays aware of stiffness changes while maintaining task focus (Appendix A).

Practical usability is another crucial factor. EMG-based impedance control methods [17]–[19] require electrode placement and frequent recalibration, making them impractical for fast-paced teleoperation. R4 ensures that the interface minimizes setup time and calibration procedures, allowing for rapid deployment and user convenience.

Lastly, impedance control interfaces must prevent unintended interactions caused by force feedback. Systems such as EMG-based approaches [17]–[19] inherently link stiffness adjustments to involuntary muscle responses, reducing control accuracy. R5 ensures that force feedback remains separate from stiffness commands, avoiding unintended impedance changes due to reflexive stiffening [10]. By maintaining independent control channels, the interface preserves precision and user intent.

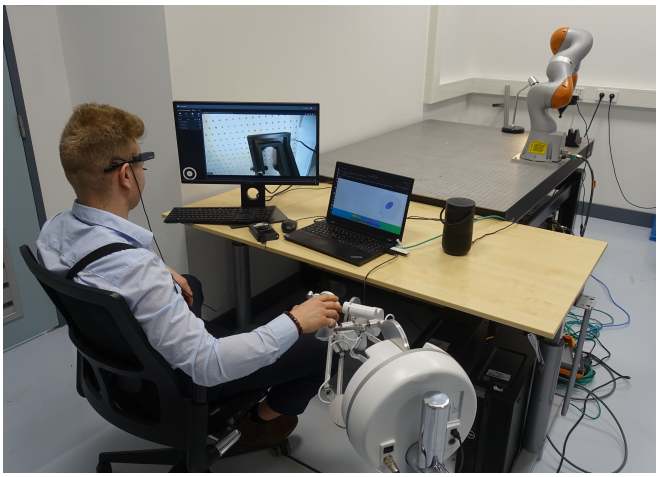

Fig. 3. Image of the teleoperation setup illustrating the operator at the local site interacting with the haptic device using their right hand, while simultaneously monitoring a real-time video feed from the remote site displayed on a monitor. The operator wears mobile eye-trackers in the form of glasses, providing gaze estimates used by the interface. The laptop runs the user interface, enabling the activation of hands-free mode through verbal commands. In this mode, the visio-verbal impedance interface is operated entirely through speech, including system calibration procedures. The remote robot arm is visible at the top right.

B. Visio-Verbal Teleimpedance Interface Design

The proposed *visio-verbal* teleimpedance command interface addresses key limitations of existing teleimpedance control methods by introducing a shared control paradigm that combines gaze and speech inputs with real-time multimodal processing. This approach simplifies impedance adjustments by eliminating the need for physical controls such as scroll wheels, joysticks, or force sensors. By using gaze and verbal commands, the interface allows operators to intuitively configure impedance parameters without manual interventions. This method meets the design requirements outlined in Section II-A as follows:

The interface combines eye-tracking with verbal interaction to generate the impedance parameters, adhering to a human-in-the-loop shared control paradigm (R1). The vision-language model (VLM) aids in defining and orienting the impedance ellipsoid while allowing the operator to refine configurations verbally. This balance effectively bridges automation with manual input, overcoming limitations found in fully automated or entirely manual control methods [24].

The system provides operators with full control over the 3D translational endpoint impedance (R2). Unlike EMG-based interfaces, which mostly scale impedance magnitude without providing precise control over orientation, the proposed system allows real-time modifications of the impedance ellipsoid's size, shape, and orientation, significantly enhancing flexibility and task-specific adaptability.

Additionally, the interface ensures operators maintain visual focus on the primary teleoperation task, avoiding distractions

caused by interacting with external graphical interfaces or additional manual controls (R3). Leveraging gaze tracking eliminates the need to shift visual attention away from the task, promoting uninterrupted workflow, particularly crucial during precise telemanipulation operations.

By relying solely on gaze tracking and speech recognition, the proposed system reduces the necessity for extensive calibration and lengthy setup procedures, providing fast deployment and immediate readiness compared to EMG-based interfaces that require complex sensor placement and frequent recalibration (R4) [17]–[19].

Lastly, the interface circumvents neuromechanical coupling effects between commanded impedance and force feedback (R5). Since the system does not depend on biomechanical signals such as muscle activation or grip force, it prevents unintended impedance alterations resulting from reflexive responses. This decoupling maintains precise impedance control and avoids unintended adjustments common in coupled interfaces [18], [19].

The proposed *visio-verbal* teleimpedance command interface addresses the key gaps identified in existing teleimpedance methods, as detailed in Section I. Current manual methods, including EMG-based interfaces, force grip sensors, and push-button systems, generally suffer from complex setup processes, limited degrees of freedom, or unintended coupling effects. On the other hand, fully automated systems, such as vision-based methods, completely remove human oversight, reducing adaptability in unpredictable environments. Additionally, current approaches frequently fail to offer intuitive, full 3D impedance control without compromising the operator's visual attention on the primary task.

To overcome these limitations, this work introduces the first *visio-verbal teleimpedance interface*, combining gaze-based contextual awareness with reasoning capabilities from a large language model (LLM) to dynamically adjust impedance. By removing the reliance on direct manual controls, the interface significantly reduces the physical and cognitive demands placed upon the operator. Simultaneously, the system maintains necessary human oversight through its shared control approach, striking an effective balance between automation and operator refinement. The interface facilitates full 3D impedance control, providing flexibility that does not depend on restrictive hardware or extensive training. Furthermore, by circumventing biomechanical coupling, the system ensures predictable and stable impedance behaviors, eliminating involuntary adjustments. Collectively, these innovations represent a significant advancement in teleimpedance control, effectively bridging the gap between manual and automated solutions while ensuring intuitive, efficient, and adaptable impedance configuration.

C. Hardware Setup

Figure 4 illustrates the hardware architecture of the teleimpedance interface. The system is divided into two main locations: the remote site (green block) and the local site, where the operator is positioned (outside of the green block).

At the local site, the operator monitors the remote environment via a display screen (right side of Figure 4), which

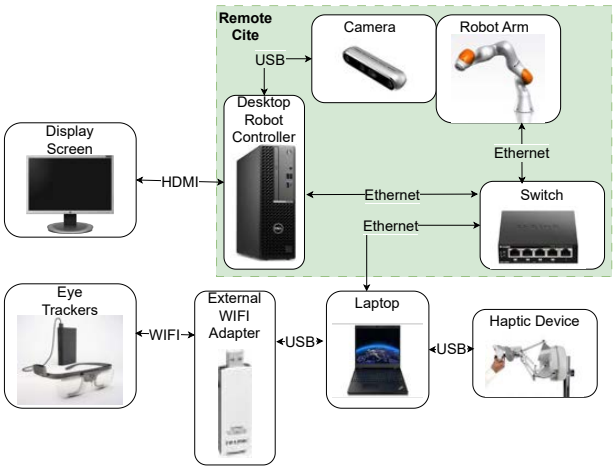

Fig. 4. Connection diagram showing hardware components and their communication links.

presents the live feed from a camera mounted directly on top of the robot's end-effector (top right). Positional commands are issued through a haptic device (bottom right), while impedance adjustments are determined based on the operator's gaze, recorded by an eye-tracking system (bottom left), and the operator's verbal commands, captured by the laptop's built-in microphone (bottom center). The operator's laptop serves as the central processing unit, receiving gaze input, processing verbal commands, and transmitting all relevant data to the remote site.

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 laptop screen. This setup ensures that impedance control is intuitive, allowing the operator to dynamically adjust stiffness parameters without disrupting task execution.

For detailed specifications of the hardware components used in this implementation, refer to Appendix G. While the hardware forms the system's physical foundation, the software architecture unifies its components, integrating multimodal inputs and outputs to enable on-demand stiffness updates and ensure seamless interaction between the operator and the remote robot to achieve a human centric interface. The next section provides an overview of the software framework that enables on demand teleimpedance control.

D. Software Architecture

One of the key strengths of the proposed *visio-verbal* teleimpedance interface is its rapid startup and modular design, made possible by a Docker-based architecture. By using `docker-compose`, the whole system is already up and running on laptop startup, provided the hardware connections are ready (see Figure 4). Each subsystem runs in its own container, isolating software dependencies and preventing version conflicts. This design minimizes setup time and technical

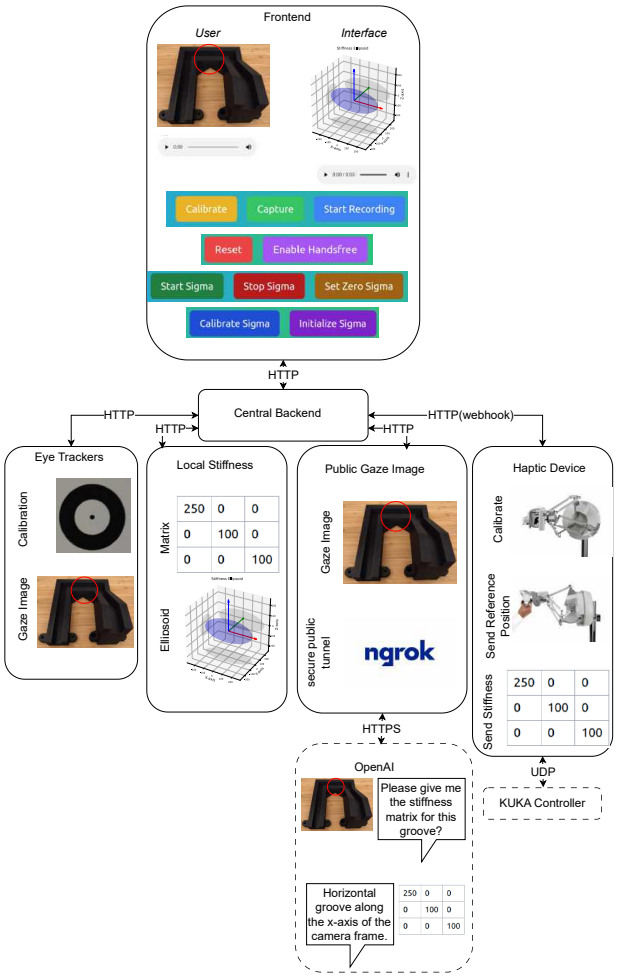

Fig. 5. Connection diagram showing software containers and their communication links. A full page version can be found in appendix A

overhead, aligning with the minimal setup requirement (Section II-A).

Before deployment, environment variables (e.g., port mappings, authentication tokens for OpenAI or Ngrok) can be adjusted to suit specific teleoperation scenarios. If the operator opts not to use an eye tracker, for instance, excluding that container from the `docker-compose` run command immediately removes its functionality without affecting other components. Using docker containers, different Ubuntu versions or library configurations can therefore coexist across containers, enabling seamless experimentation and simple subsystem swaps.

The Eye Tracker Container handles instant eye tracker calibration, captures gaze data, and overlays gaze estimates on snapshots. Its hardware dependencies require that eye trackers are powered on and connected to the laptop via an external Wi-Fi adapter. Similarly, the Haptic Interface Container manages the calibration procedure for the haptic device and coordinates UDP transmissions to the robot controller, including position commands and real-time stiffness updates. It provides start/stop control for UDP messages, preventing unintended robot movements when the operator is away, and includes a *set zero* feature to align the haptic device's zero position with the robot's position, avoiding sudden positional jumps. The haptic interface must be plugged in, powered on, and connected to the laptop via USB.

The Local Static Server stores and serves generated stiffness matrices and ellipsoids, making them accessible to both the frontend and central backend. The frontend retrieves these ellipsoids to visually inform the operator of the current impedance configuration, while the central backend utilizes stiffness matrices to maintain and contextualize the conversation history with the vision-language model (VLM). This server requires the operator's laptop to be powered on.

The Public Static Server stores and serves snapshots of the remote site, which are overlaid with gaze estimates represented by red circles. It establishes an internet-accessible tunnel via ngrok, enabling the online vision-language model (VLM) to retrieve gaze images through web-accessible URLs. This server also requires the operator's laptop to remain powered on.

The Central Backend routes requests from the frontend to relevant containers (eye tracker, haptic interface, local/public servers), manages conversation history, and interfaces with the VLM for stiffness matrix generation. It also provides speech-to-text and text-to-speech capabilities, and similarly depends on the laptop being powered on.

Finally, the Frontend Container delivers the user interface, including controls for calibration, image capture, audio recording, and haptic device status, with the primary hardware requirement again being that the laptop is powered on.

By containerizing each subsystem, operators can effortlessly adapt the system to various teleoperation tasks while avoiding software conflicts (e.g., Python or Ubuntu version mismatches). Rather than juggling multiple virtual environments, `docker-compose` launches every component with its specific software stack, thereby reducing maintenance time and complexity. The next section zooms into the VLM interaction as this requires additional explanation.

E. VLM Interaction: Prompt Engineering

The rapid advancements in Deep Learning, particularly the transformer architecture [27] and the increasingly larger model sizes [28], have enabled breakthrough applications in speech-to-text (STT), text-to-speech (TTS), and natural language processing combined with computer vision through vision-language models (VLMs). A popular example is ChatGPT, which uses a VLM to process both textual and visual inputs [29]. While these models 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 AI's context-awareness by identifying where a user is looking [30]. Building on their findings, integrating eye tracking into a visio-verbal impedance system emerged as a natural progression to attempt VLM driven teleimpedance control.

Vision-language models (VLMs) extend the capabilities of large language models (LLMs) by incorporating image

processing alongside textual input. Research on LLMs has demonstrated that increasing model size leads to improved performance [28]. A key discovery with these larger models was their ability to perform few-shot learning, where instead of fine-tuning a model for a specific task, researchers provide a small set of example question-answer pairs alongside the user's prompt. This approach enables the model to generalize the task and its corresponding labels more effectively than fine-tuning on a dedicated dataset [31].

The introduction of VLMs naturally extended this few-shot learning paradigm to multimodal tasks, where models benefit from contextual examples in both text and vision-based applications. 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 [32]–[34]. This research suggests a natural extension of few-shot learning in VLMs by incorporating gaze-estimate images alongside a curated set of labeled examples that define the corresponding stiffness matrices. This approach leverages the few-shot paradigm to enhance the model's ability to infer appropriate impedance configurations based on visual and contextual cues.

Figure 6 illustrates the structure of the prompt used for communication with the vision-language model (VLM). The prompt comprises three core components. First, the system role defines the task by instructing the VLM to operate as a specialized "stiffness matrix generator." This ensures the user's input is consistently interpreted in the context of stiffness matrix updates rather than general dialogue. The term "system role" originates from deep learning terminology and can be understood as the task description. These terms are used interchangeably throughout the paper. Second, an optional few-shot demonstration can be provided. This consists of a concise set of example queries along with their corresponding outputs, effectively acting as a "lookup table" to guide and enhance the model's task-specific response capability by enabling it to infer patterns from prior demonstrations. Finally, the conversation history component maintains a record of the operator's past commands and contextual information, including snapshots with gaze estimates. This ongoing history allows for iterative refinement of the stiffness matrices, preventing the model from resetting after each interaction and thus ensuring continuity and consistency in impedance adjustments.

Additionally, the vision-language model's (VLM) image-processing detail level significantly influences system performance, and two modes are available to balance accuracy and responsiveness. The low-detail mode downscales images to 512×512 pixels and allocates 85 tokens for their description. This approach reduces computational overhead and accelerates response time, though at the expense of finer-grained feature detection. Conversely, the high-detail mode provides more detailed image analysis by allocating additional tokens (approximately 170 per image tile). This enhances the model's ability to accurately extract gaze-based visual features, but also introduces increased latency and token consumption.

These prompt components collectively define how the VLM processes gaze and verbal input to generate an appropriate stiffness matrix, allowing for a trade-off between computational efficiency and interpretability based on task requirements.

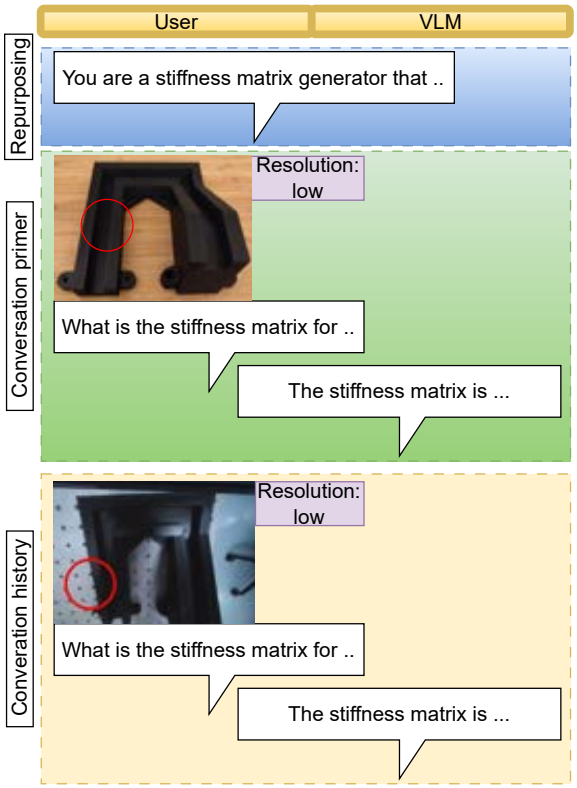

Fig. 6. Illustration of the multi-part prompt, which consists of repurposing the general VLM with a task-specific system role (blue), an optional conversation primer containing examples of desired behavior for few-shot learning (green), and an ongoing conversation history (yellow). Maintaining this conversation history enables continuity in interactions and refinement of responses. Image resolution is also highlighted which can be set to high or low depending on the level of detail needed to perform the vision task.

The next section discusses the experiments and their results.

III. Experiments

To validate and optimize the proposed *visio-verbal* teleimpedance interface, two experiments were conducted. Experiment 1 involved prompt optimization to identify optimal parameters for effectively prompting the proprietary vision-language model (VLM), GPT-4o from OpenAI. [29] Parameters included the complexity of the task description, ranging from minimal instructions to elaborate context with explicit stiffness matrix labels, the presence or absence of prior examples, and image detail levels, all significantly influencing the VLM's accuracy. Experiment 2 consisted of two parts: Experiment 2a established a baseline by demonstrating the interface's functionality using only verbal commands, without gaze-based inputs, while Experiment 2b tested the selected optimal prompt configuration within the complete teleimpedance setup to evaluate practical feasibility and real-time performance. Both experiment 2a and 2b utilized the same 3D-printed groove

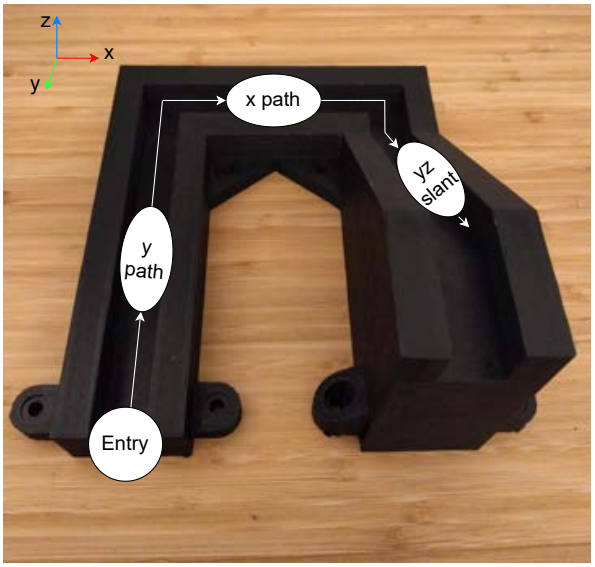

Fig. 7. Groove structure detailing each section for which the impedance interface predicts stiffness configurations. The two-dimensional ellipses indicate the desired stiffness shapes and orientations for each groove section in the x-y plane. Because stiffness configurations are three-dimensional, these ellipses provide only partial representations, excluding the z-axis component. For a complete visualization of the 3D stiffness configurations, ellipsoids are used, as illustrated in Figure 8. All corresponding 3D stiffness ellipsoids for each groove section are detailed in Appendix H.

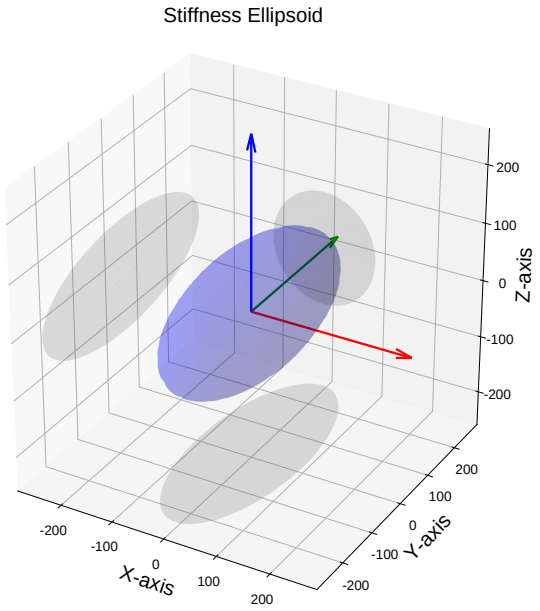

Fig. 8. An example of a stiffness ellipsoid derived from a VLM-generated stiffness matrix, providing a graphical representation of the robot's endpoint translational stiffness. Larger radii correspond to higher stiffness, while smaller radii indicate lower stiffness. This particular ellipsoid represents an endpoint stiffness matrix of $\mathrm{diag}(100, , 250, , 100)$, corresponding to a groove aligned with the y-axis, along which the robot endpoint moves. Stiffness is highest along the y-axis (elongated shape), ensuring good position tracking in that direction, while lower stiffness along the x- and z-axes (narrower shape) minimizes high unwanted interaction forces with the groove to minimize friction and to prevent potential damage.

structure to systematically assess how effectively the interface could generate and adjust impedance parameters in a realistic slide-in-the-groove teleoperation scenario.

A. Motivation: The Slide-in-the-Groove Task

Consider the groove structure illustrated in Figure 7, designed specifically to evaluate variable impedance control capabilities of the proposed interface. Throughout the task, impedance properties must be dynamically adapted to the groove's distinct sections. At the groove's entrance, the robot arm is carefully lowered and aligned without exerting excessive force. This gentle alignment is facilitated by setting a *low and equal stiffness* in the x and y directions, allowing the peg to naturally self-align within the horizontal plane, while assigning a *high stiffness* in the z direction to maintain precise vertical position tracking. Once fully inserted into the groove, stiffness in the *forward direction* is increased to ensure accurate tracking of the reference trajectory, whereas stiffness in the *lateral directions* remains low to allow compliance with groove walls. This configuration prevents excessive normal forces, reducing risks of jamming or structural damage. An example of the full 3D stiffness configuration, depicted as a 3D stiffness ellipsoid for traversal along the y-direction, is provided in Figure 8. Appendix B-B provides the derivation of the ellipsoid, along with the corresponding ellipsoids for all other stages of the groove structure, and includes their respective plots.

This explanation illustrates why variable stiffness control is essential for effectively traversing the groove structure and successfully completing the slide-in-the-groove task. The primary evaluation criterion for the feasibility of the proposed *visio-verbal* impedance interface is the capability of the impedance interface to generate stiffness matrices matching those illustrated in Figure 7 for each specific section of the groove structure.

B. Experiment 1: Prompt Optimization

As discussed in Section II-E, the way the VLM is prompted significantly influences the quality of its outputs. To ensure optimal performance, we conduct a parameter optimization experiment to identify the most effective combination of prompt parameters, including the task description, example demonstrations, and image detail.

The experiment aimed to identify the optimal combination of parameters for generating accurate stiffness matrices. Three distinct system roles were tested, each progressively including more information. First, a minimal task description was provided, offering only basic context. Second, an elaborate task description was tested, adding extra details to clarify the task. Finally, the third approach combined the elaborate task description with explicit labels, presenting example stiffness matrices to further guide the model. The complete descriptions of these roles can be found in Appendix F.

Beyond the system roles, another critical factor evaluated was the inclusion of a prior message list containing ex-

ample prompts and corresponding desired stiffness matrices. As discussed in Section II-E, large language models are inherently few-shot learners and typically demonstrate improved performance when provided with representative task examples. [31] Thus, the experiment investigated whether this principle also applies to a vision-language model tasked with generating accurate stiffness matrices in a teleimpedance context. Three conditions were tested. The first condition excluded a prior message list, meaning the model operated without any examples. The second condition incorporated a prior list derived from an ideal environment to examine the impact of environmental variability on model performance. The third condition used a prior list from the lab environment to determine whether familiarity with the specific operational context improved stiffness matrix accuracy.

Lastly, the image detail was adjusted between low and high settings to assess its impact on performance. The low-detail mode resulted in faster model responses, while the high-detail mode allowed the extraction of finer image features at the cost of increased computational demand and processing time, as detailed in Section II-E. Combining the three task description variants, three prior message list conditions, and two image detail levels resulted in 18 parameter combinations, systematically evaluated to determine their effect on the accuracy of the generated stiffness matrices.

Since the groove structure consists of four distinct stages, each parameter combination required four predictions, one per stage, to evaluate its performance throughout the groove structure. However, as discussed in Section II-E, vision-language models exhibit stochastic behavior, meaning that repeated runs with identical inputs may yield different outputs. To account for this variability, each combination was tested five times per stage.

Performance was quantified using an accuracy metric. Even minor deviations from the correct stiffness configuration would necessitate corrective intervention from the operator via additional voice commands, making numerical proximity less relevant. Subsequently, to establish statistical significance, a 95% confidence interval analysis identified the top-performing combinations. Those combinations whose confidence intervals overlapped with the most accurate configuration underwent a second evaluation, comprising 20 predictions per stage. This expanded evaluation narrowed the confidence intervals, thereby enhancing the statistical robustness and reliability of the experimental results.

1) Prompt Optimization - Results: Figure 9 illustrates the confidence intervals for all 18 tested parameter combinations. Each combination was evaluated across four test images, corresponding to the four groove structure stages, with five trials per image, resulting in a total of 20 trials per combination. The accuracy results show variability across different configurations. While the highest-performing combination cannot be statistically distinguished from several others at a 95% confidence level (p-value = 0.05), 10 of the 18 combinations exhibit significantly lower accuracy and fall outside the confidence interval of the top-performing configuration, highlighted in red in Figure 9. These lower-performing combinations were

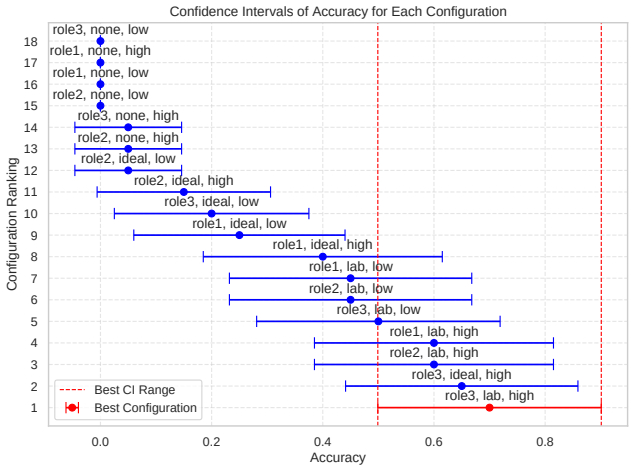

Fig. 9. Confidence intervals of accuracy for the 18 tested prompt configurations. Each interval represents a specific parameter combination, with system role, prior, and image detail indicated above the corresponding interval. The red configuration highlights the best-performing combination. All configurations that fall within the red interval, as indicated by the red dotted line, were selected for further evaluation to establish a more statistically significant distinction between them.

excluded from further evaluation, narrowing the selection to the top eight statistically competitive configurations.

In the second evaluation, the eight best-performing parameter combinations were further tested by predicting the stiffness matrix for each groove stage 20 times. The resulting rankings are presented in Table I. The last column of the table presents accuracy scores excluding the slant stage, as all combinations performed poorly on this segment. This is particularly relevant, as discussed in Section IV, where we examine this result in greater depth. We propose a hypothesis to explain the poor performance on the slant, evaluate supporting evidence for its validity, and discuss its implications for assessing the feasibility of the proposed impedance interface.

At the conclusion of this experiment, we determined the optimal prompt configuration: an elaborate task description with labeled examples for the system role, high-detail image processing, and the inclusion of lab-based exemplars as a primer. This refined configuration was then applied in the second phase of experimentation, where the interface was tested in a real-time teleimpedance setup.

C. Experient 2: Demonstration of the Visio-Verbal Teleimpedance Interface

After determining the optimal prompt configuration, it was integrated into the visio-verbal teleimpedance interface. Before demonstrating the full setup with this integration, an initial baseline experiment was conducted using a verbal-only interface, temporarily excluding the eye tracker. This speech-only demonstration served as both a minimal working example and a baseline for comparison while also showcasing the modularity of the interface, which allows seamless transitions between visio-verbal and verbal-only modes.

1) Verbal-only Impedance Interface: In this configuration, the operator interacts with the system exclusively through

Role	Prior	Detail	Entrance	Y-traverse	X-traverse	Slant	Overall Accuracy	Overall Accuracy (No Slant)
role3	lab	high	1.00 ± 0.00	0.93 ± 0.06	1.00 ± 0.00	0.00 ± 0.00	0.73 ± 0.06	0.98 ± 0.02
role1	lab	high	1.00 ± 0.00	0.67 ± 0.12	0.93 ± 0.06	0.13 ± 0.09	0.68 ± 0.06	0.87 ± 0.05
role2	lab	high	0.93 ± 0.06	0.67 ± 0.12	1.00 ± 0.00	0.07 ± 0.06	0.67 ± 0.06	0.87 ± 0.05
role3	ideal	high	0.93 ± 0.06	0.67 ± 0.12	0.93 ± 0.06	0.00 ± 0.00	0.63 ± 0.06	0.84 ± 0.05
role1	lab	low	1.00 ± 0.00	0.27 ± 0.11	1.00 ± 0.00	0.00 ± 0.00	0.57 ± 0.06	0.76 ± 0.06
role2	lab	low	1.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00	0.50 ± 0.06	0.67 ± 0.07
role3	lab	low	1.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00	0.50 ± 0.06	0.67 ± 0.07
role1	ideal	high	0.73 ± 0.11	0.40 ± 0.13	0.60 ± 0.13	0.00 ± 0.00	0.43 ± 0.06	0.58 ± 0.07

TABLE I

SECONDARY EVALUATION OF THE 8 BEST-PERFORMING PARAMETER COMBINATIONS. THE VALUES REPRESENT ACCURACY WITH STANDARD DEVIATIONS (MEAN ± STD, ROUNDED TO TWO DECIMALS). ACCURACY FOR EACH INDIVIDUAL SECTION IS CALCULATED OVER 20 TRIALS, WHILE THE OVERALL ACCURACY IS DETERMINED FROM ALL 80 TRIALS COMBINED. THE LAST COLUMN REPORTS OVERALL ACCURACY EXCLUDING THE SLANT SECTION, THEREFORE BASED ON 60 TRIALS.

voice commands, without providing gaze snapshots. Without visual input, the operator must provide more detailed verbal instructions, explicitly specifying the groove section or orientation that would otherwise be inferred from the image. The model infers the appropriate stiffness matrix based on a custom system role designed for the VLM, which provides a structured task description with predefined section labels to guide its predictions. For example, a label such as "A groove along the X-axis" corresponds to the diagonal stiffness matrix

$$K = \begin{bmatrix} 250 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 100 \end{bmatrix} \quad (1)$$

The user can leverage this structure by stating, for instance, "I need the stiffness matrix for a groove along the X-axis."

Since no images are involved in this setup, it did not include an image list prior or a detail parameter. This demonstration validates the interface's functionality without visual input, highlighting its adaptability by relying solely on explicit user-defined axis and stage names instead of gaze-based visual context.

2) Verbal-only Impedance Interface - Results: The results of this experiment are presented in Figure 10. The top section of the figure illustrates the groove structure, highlighting the specific section for which the interface must predict the stiffness matrix. Beginning at the entrance stage, the operator issues a voice command such as, "Please provide the stiffness matrix for the entrance." In response, the interface sets the stiffness to be high along the z-axis at 250 N/m while maintaining compliance in the x- and y-axes with a stiffness of 100 N/m. Using similar voice commands, the operator sequentially specifies the stiffness matrix for subsequent stages, progressing through the y-path, x-path, and slanted section.

The bottom plot of Figure 10 confirms that the impedance was adjusted correctly for each section of the groove structure, demonstrating the system's ability to apply task-specific stiffness values as instructed. The second plot from the bottom indicates that the highest interaction force recorded during the task was approximately 10 N, which is well within the acceptable range for this application [22].

The top two plots depict the reference position and the measured position. Discrepancies between these plots typically occur when the reference position lies within the groove structure's wall, preventing the peg from reaching it. In such cases,

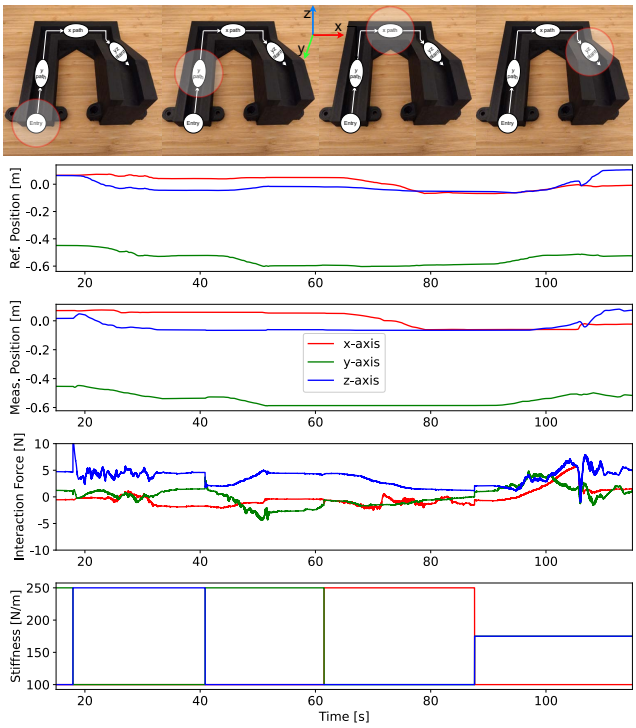

Fig. 10. Variable impedance control for the *audio only* setup through the groove structure, including entrance, ytraverse1, xtraverse and the slant.

the interaction force between the peg and the wall increases. However, with the appropriate stiffness matrix applied, these forces remain within the acceptable threshold of 10 N.

3) Visio-verbal Impedance Interface: Following the successful demonstration of the verbal impedance interface in the previous section, the next step is to integrate the optimal parameters identified in Experiment 1, as described in Section III-B, and evaluate the visio-verbal impedance interface. In this setup, the operator's gaze estimate, overlaid on a remote scene snapshot, is incorporated into the VLM prompt to provide stiffness matrix prediction. The operator selects a section of the groove structure by directing their gaze and capturing a snapshot. They then issue a stiffness update request (e.g., "Please provide the stiffness for this part of the groove structure"), enabling the VLM to utilize both the voice transcription and

the remote scene snapshot with the gaze estimate to generate the corresponding stiffness matrix.

During the demonstration, the operator traverses from the entrance stage through the y-path, concluding at the x-path. The slanted section is excluded due to the suboptimal performance of the optimal configuration in this region, as indicated in Table I. Further discussion on this limitation is provided in Section IV. Upon reaching the x-path, the demonstration highlights the memory capabilities of the interface by guiding the operator back through the groove structure along the same path. Without capturing additional snapshots, the operator retrieves previously generated stiffness matrices and applies them to the robot using commands such as, "I would like to move back through the groove structure, please apply the previous stiffness matrix."

4) Visio-verbal Impedance Interface - Results: Figure 11 presents the results of the visio-verbal impedance interface demonstration. Similar to the verbal-only experiment, the top section of the figure highlights the selected sections of the groove structure as the operator progresses from the entrance to the x-path and back. Below this, the issued prompts are displayed, where the first three prompts include gaze images, while the last two exclude them to emphasize the system's memory capability.

The bottom plot of the figure shows a symmetric stiffness progression, which already suggests a desirable outcome, as the operator follows the same path in both directions. Notably, the last two stiffness adjustments, made while leveraging the system's memory, correctly matched the previously set stiffness values, demonstrating the interface's ability to recall and reapply prior settings without requiring additional visual input.

The lowest plot in the figure represents the stiffness settings applied to the robot during the demonstration. Initially, the system sets a high stiffness of 250 N/m along the z-axis while maintaining low stiffness (100 N/m) in the x- and y-directions. This configuration is ideal for the entrance stage, where the operator needs to lower the peg into the groove, requiring precise position tracking in the z-direction. As the operator moves along the y-path, the system adjusts to prioritize stiffness along the y-axis to ensure stable movement while maintaining compliance in the other directions. Similarly, upon reaching the x-path, the system shifts its primary stiffness direction to the x-axis. When retracing the path, the system correctly restores the previously applied stiffness settings, ensuring smooth traversal back through the y-path and ultimately increasing stiffness in the z-axis again to facilitate the peg's exit from the groove.

Additionally, the interaction forces remain below 10 N, which falls within the acceptable range for this task [22]. The close alignment between the reference and measured positions further supports this result, indicating that the reference position remained outside the groove walls for most of the task, minimizing unintended contact forces.

The next section analyzes the results presented here, examining the interface's contributions, strengths, and limitations in the context of existing teleimpedance interfaces.

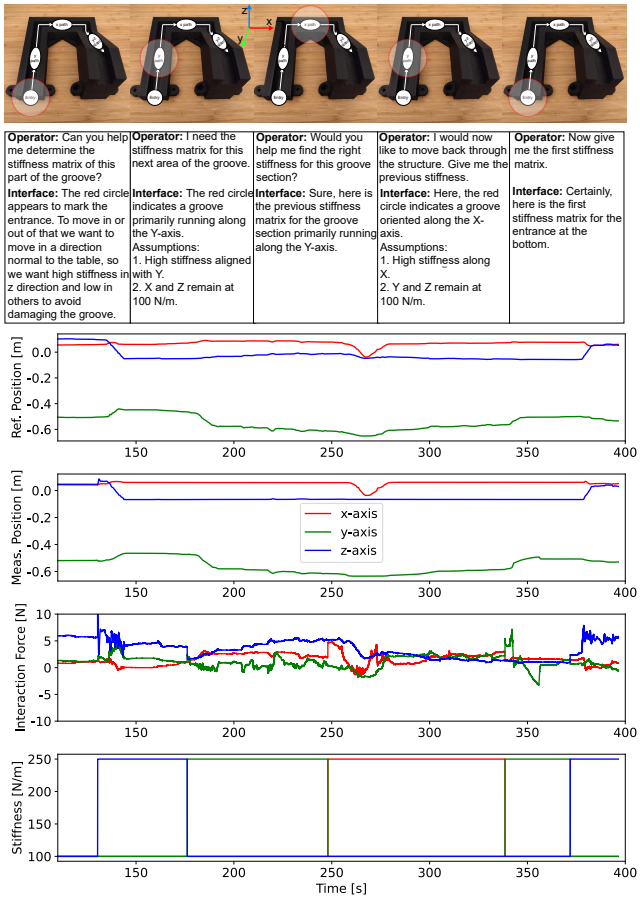

Fig. 11. Variable impedance control through the groove structure, including entrance, ytraverse1 and xtraverse

IV. Discussion

This section provides an in-depth analysis of the results presented in the previous section, beginning with the prompt optimization experiment, focusing on, but not limited to, the best-performing combination of prompt parameters. It then examines the performance of the verbal-only impedance interface, assessing its effectiveness and limitations in the absence of visual input. Finally, the discussion focuses on the full visio-verbal impedance interface, analyzing its ability to integrate gaze-based context and memory recall in stiffness prediction for the slide-in-the-groove task.

Key aspects explored in this section include the implications of stiffness matrix accuracy, the effectiveness of gaze input in reducing verbal instruction complexity, and the role of memory in enhancing task continuity. Additionally, potential sources of error, system limitations, and areas for further optimization are discussed.

A. Prompt Optimization

The results in table I indicate that the top four configurations all utilize high image detail. This suggests that the VLM performs better with the high-detail setting, as described in Section II-E. Next, table I also shows that the best 3

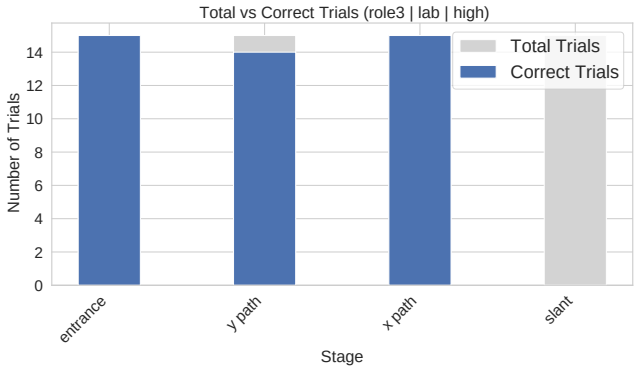

Fig. 12. Plot of the trials of the best-performing prompt configuration: Role 3, lab prior, and high image detail. This combination achieved near-perfect accuracy, with only one error across 45 trials spanning the entrance, y-path, and x-path sections.

performing combinations utilize the lab prior in which the examples contain images made with the lab setup on which the prompt combinations were also tested. This indicates that one can get a performance boost by making sure that the prior message list contains image examples that are very similar to the images on which the model needs to perform the task. Although the ideal prior message list contains images with better lighting the fact that it was different from the test images caused it to perform worse. Finally there is a significant difference between role 3 with respect to role 1 and 2 emphasizing the importance of including a list of labels in the task description of the VLM when subjecting it to this task. This list of class labels including the 4 different stages of the groove structure is therefore considered as a performance booster. Figure 12 illustrates the performance of the best-performing configuration, which utilizes Role 3 as the VLM's task description, the lab prior message list, and high image detail. The figure demonstrates near-perfect accuracy across the first three stages of the groove structure. However, like most other configurations (as detailed in Appendix D), it fails entirely in the slant section, achieving 0

Examining this behavior further, it was observed that for the best-performing combination, Role 3, lab prior, and high image detail, the VLM consistently predicted the stiffness matrix corresponding to the y-path in 100% of trials. A visual inspection of the prior message list images, as shown in Appendix E, reveals a likely cause for this outcome. The camera angle, determined by the endpoint-mounted camera, captures the groove structure from a nearly top-down perspective. Figure 13 provides an example of how the slant section appears from this viewpoint.

To validate the hypothesis that the camera angle contributed to the poor accuracy for the slant section, the best-performing combination was retested using the slant image from the ideal setup. While this angle still lacks optimal height perspective, as shown in Figure 14, it significantly improved accuracy to 66.66%, supporting the hypothesis that the camera angle affected the model's ability to recognize the slant.

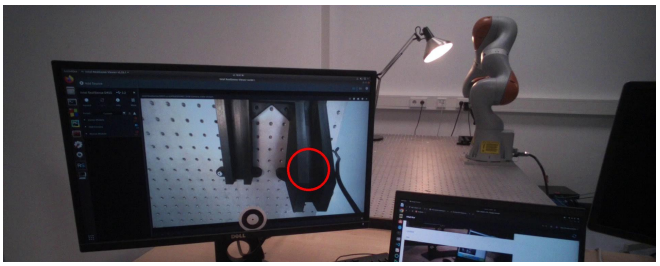

Fig. 13. Image of the slant section captured by the robot's endpoint-mounted camera, illustrating the top-down camera angle. This perspective limits depth cues, making it difficult for the VLM to distinguish the slant from the y path.

B. *Visio-verbal Impedance Interface*

The results of the verbal-only impedance interface experiment demonstrate that the system effectively applies task-specific stiffness values based solely on spoken commands. The successful impedance adjustments at each stage of the groove structure, as shown in Figure 10, indicate that the interface can operate without gaze-based input while still generating appropriate stiffness matrices.

The interaction forces remained within acceptable limits, with a peak force of approximately 10 N, confirming that the selected stiffness configurations allowed for smooth traversal through the groove structure. The minimal force deviations observed suggest that the interface provided sufficient compliance to prevent excessive interaction forces.

Despite these positive results, the verbal-only approach has certain limitations. The operator must explicitly specify the groove section or its orientation for each command, adding cognitive load compared to the visio-verbal interface, where gaze input inherently provides this contextual information. This eliminates the need for verbal orientation descriptions, making interactions more fluid. Additionally, while the system accurately predicted stiffness values for all sections, the lack of visual cues may hinder adaptability in more complex or dynamic environments where the operator may struggle to accurately judge the groove orientation. These findings confirm the verbal-only interface as a viable fallback option while underscoring the advantages of integrating gaze input for a more intuitive and efficient teleimpedance control experience.

The results of the visio-verbal impedance interface demonstration highlight the benefits of integrating gaze-based input with verbal commands for teleimpedance control. Compared to the verbal-only interface, the visio-verbal system reduces cognitive load specifically for providing context regarding the groove orientation by allowing the operator to specify stiffness adjustments more intuitively through gaze, without the need for explicit section naming or orientation naming. The ability to leverage gaze images provides additional contextual information, enabling the model to infer the correct stiffness configuration more efficiently without verbal help from the operator.

Another key strength of the interface is its memory capability, as demonstrated in the last two stiffness adjustments, where previously set values were recalled and applied without requiring new gaze inputs. This feature streamlines interaction

Fig. 14. Image of the slant section captured from a more favorable camera angle. This perspective enhances depth cues, improving the VLM's ability to differentiate the slant from the y-path.

by reducing redundant commands, allowing for a more fluid and efficient teleoperation experience. The symmetric stiffness progression shown in Figure 11 further validates the system's consistency, as stiffness settings were correctly restored when retracing the path. Additionally, the return path was completed in less time than the initial traversal, indicating improved efficiency and adaptability in repeated operations.

Additionally, the interface maintained interaction forces below 10 N, ensuring safe and stable operation throughout the task. The close match between reference and measured positions suggests that the system successfully adjusted stiffness to avoid unintended contact forces while maintaining trajectory accuracy. These results confirm that the visio-verbal impedance interface offers an effective and intuitive solution for on-demand teleimpedance control, improving both usability and task efficiency compared to a speech-only approach.

Both experiments validate that carefully designed prompts enable the proprietary VLM to generate appropriate stiffness matrices for the slide-in-the-groove task. Experiment 1 identified an effective prompt configuration, while Experiment 2 evaluated the interface's practical performance in real-time teleimpedance control. Together, these findings demonstrate that the *visio-verbal* interface can dynamically generate adaptive stiffness configurations in response to user commands and visual input, addressing the need for intuitive interaction in unstructured robotic environments.

Revisiting the requirements outlined in Section II, which address gaps in current teleimpedance interfaces, the results confirm that the visio-verbal teleimpedance interface successfully meets all specified criteria, demonstrating its feasibility. By fulfilling these requirements, the interface also establishes its relevance within the broader context of existing teleimpedance research.

C. Limitations

The most evident limitation is the interface's poor performance on the slant section of the groove structure. As previously discussed, this issue is likely due to the top-down camera view, which lacks the necessary visual cues for height perception. However, further testing with the ideal setup image demonstrated that a more favorable camera angle significantly improves performance. This suggests that, given sufficient visual cues, the interface could achieve accuracy comparable to that of other sections.

Another key limitation of this study is the reliance on a mobile eye tracker, which requires additional image-processing steps to isolate the display screen from its surroundings. While an automatic cropping function was implemented to mitigate this issue, its performance was inconsistent. In some instances, it successfully captured the 3D-printed structure and gaze estimate, while in others, it excluded both entirely. The final iteration of the interface therefore excluded the cropping functionality and used the uncropped images as taken by the eye trackers.

Additionally, streaming was not incorporated into the pipeline. Although the VLM generates responses as a stream of words, each word could theoretically be sent directly to the text-to-speech algorithm upon arrival. However, given the brevity of the verbal notifications, this approach is unlikely to significantly reduce response time. Consequently, investing substantial effort in implementing streaming is not recommended. While streaming can enhance response time in systems that generate long textual outputs, it was deemed unnecessary for this teleimpedance interface. The VLM was explicitly prompted to produce concise responses, consisting mainly of a stiffness matrix (filtered out before speech synthesis) and a short confirmation message. Since the text-to-speech system processes only a brief, predefined phrase rather than a long, dynamically generated response, the benefits of streaming would be negligible.

Finally, the interface does not account for adaptive rotational stiffness. The slide-in-the-groove task primarily relies on translational compliance to ensure smooth insertion and movement along the constrained path, with no peg rotations required. By maintaining a fixed rotational stiffness, the system simplifies computation and interaction, allowing full focus on translational adjustments. However, in tasks involving rotational movements, extending the interface to accommodate rotational stiffness adjustments would be a necessary step for broader applicability.

The next section outlines the key findings, explores the broader impact of this research, and proposes directions for future work.

V. Conclusion

The visio-verbal teleimpedance command interface presented in this study advances the field of teleimpedance by introducing an intuitive, shared-control paradigm that balances automation with human intent. By integrating gaze and speech as natural input modalities, the interface allows operators to adjust 3D translational stiffness in real-time without diverting visual attention from the teleoperation task. This hands-free control method reduces cognitive load and eliminates the need for additional physical input devices, making stiffness modulation more seamless and accessible.

The experiments validate the feasibility of this approach. The system was capable of generating task-appropriate stiff-

ness matrices with high accuracy, demonstrating its effectiveness in an experimental slide-in-the-groove task. Conversation history management further improved usability by enabling iterative refinement and memory-based stiffness adjustments, reducing the need for redundant commands. Additionally, the interface maintained stable interaction forces within acceptable limits, reinforcing its suitability for real-world teleoperation scenarios.

The interface was validated through a 3D test groove structure, demonstrating its functionality and robustness in a realistic teleoperation scenario. This successful demonstration highlights the potential of combining visio-verbal interaction with teleimpedance control for intuitive and task-specific robot command.

While the interface performs well across most groove sections, challenges remain in recognizing slanted surfaces due to limitations in depth perception from the current camera setup. Future research should focus on optimizing camera placement or integrating depth-sensing technologies to address these limitations. Moreover, further studies should compare this interface against traditional teleimpedance methods through systematic user evaluations to assess its practical advantages in real-world applications.

Beyond teleimpedance, this approach could have broader implications in human-robot collaboration, particularly in scenarios requiring precise stiffness adaptation, such as robotic assembly, minimally invasive surgery, or remote maintenance tasks. By leveraging vision-language models and multimodal inputs, this interface represents a step toward more adaptive, user-friendly robotic impedance control systems.

A. Future Work

As just mentioned, future research should focus on optimizing camera placement to enhance depth perception and improve the interface's performance on slanted surfaces. Exploring alternative camera mounting positions or integrating depth-sensing technology could provide the VLM with richer visual cues, addressing the limitations observed in the top-down perspective.

Additionally, systematic user studies should be conducted to assess the interface's ability to configure 3D stiffness during teleoperation compared to previously proposed methods. Such studies will offer valuable insights into its usability, effectiveness, and potential impact in real-world telerobotics applications. Further research could also extend the visio-verbal interface to adaptive control of rotational stiffness, expanding its capabilities beyond translational impedance regulation.

Rather than refining the preprocessing steps for cropping images from the mobile eye tracker, a more effective solution would be to utilize a static eye tracker, eliminating the need for extensive cropping and minimizing potential errors.

Finally, future work could explore the integration of streaming in scenarios where real-time feedback is more critical or where VLM-generated responses are more extensive, ensuring faster and more responsive system interactions.

REFERENCES

[1] R. R. Murphy, K. L. Dreger, S. Newsome, J. Rodocker, E. Steimle, and S. Tadokoro, "Use of remotely operated marine vehicles at minamisanriku and rikuzentakata japan for disaster recovery," in *Proceedings of the IEEE International Symposium on Safety, Security, and Rescue Robotics*, 2011, pp. 19–25. [Online]. Available: https://ieeexplore.ieee.org/document/6106798

[2] A. M. Okamura, "Methods for haptic feedback in teleoperated robot-assisted surgery," *Industrial Robot: International Journal*, vol. 31, no. 6, pp. 499–508, 2004. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/01439910410566789/full/html

[3] R. H. Taylor, A. Menciassi, G. Fichtinger, P. Fiorini, and P. Dario, "Medical robotics and computer-integrated surgery," in *Springer Handbook of Robotics*, B. Siciliano and O. Khatib, Eds. Berlin, Germany: Springer, 2016, pp. 1657–1684. [Online]. Available: https://link.springer.com/referenceworkentry/10.1007/978-3-319-32552-1_62

[4] B. S. Peters, P. R. Armijo, C. Krause, S. A. Choudhury, and D. Oleynikov, "Review of emerging surgical robotic technology," *Surgical Endoscopy*, vol. 32, no. 4, pp. 1636–1655, 2018. [Online]. Available: https://link.springer.com/article/10.1007/s00464-018-6079-2

[5] T. B. Sheridan, "Human–robot interaction: Status and challenges," *Human Factors*, vol. 58, no. 4, pp. 525–532, 2016. [Online]. Available: https://journals.sagepub.com/doi/10.1177/0018720816644364

[6] C. Preusche, D. Reintsema, K. Landzettel, and G. Hirzinger, "Robotics component verification on ISS ROKVISS – preliminary results for telepresence," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006, pp. 4595–4601.

[7] J. Artigas, R. Balachandran, C. Riecke, M. Stelzer, B. Weber, J.-H. Ryu, and A. Albu-Schäffer, "KONTUR-2: Force-feedback teleoperation from the international space station," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2016, pp. 1166–1173.

[8] D. R. Yoerger and J.-J. E. Slotine, "Supervisory control architecture for underwater teleoperation," in *Proceedings of the IEEE International Conference on Robotics and Automation*, vol. 4, 1987, pp. 2068–2073. [Online]. Available: https://ieeexplore.ieee.org/document/1087890

[9] J. C. Jurmain, A. J. Blancero, J. A. Geiling, A. Bennett, C. Jones, J. Berkley, M. Vollenweider, M. Minsky, J. C. Bowersox, and J. M. Rosen, "HazBot: Development of a telemanipulator robot with haptics for emergency response," *American Journal of Disaster Medicine*, vol. 3, no. 2, pp. 87–97, 2008. [Online]. Available: https://wmpllc.org/ojs/index.php/ajdm/article/view/1876

[10] L. Peternel and A. Ajoudani, "After a decade of teleimpedance: A survey," *IEEE Transactions on Human-Machine Systems*, vol. 53, no. 2, pp. 401–416, 2023, document Version: Final published version.

[11] R. Bischoff, J. Kurth, G. Schreiber, R. Koeppe, A. Albu-Schäffer, A. Beyer, O. Eiberger, S. Haddadin, A. Stemmer, G. Grunwald, and G. Hirzinger, "The KUKA-DLR lightweight robot arm – a new reference platform for robotics research and manufacturing," in *Proceedings of the 41st International Symposium on Robotics (ISR) and 6th German Conference on Robotics (ROBOTIK)*, Munich, Germany, June 2010, pp. 1–8. [Online]. Available: https://ieeexplore.ieee.org/document/5756872

[12] S. Haddadin, S. Parusel, L. Johannsmeier, S. Golz, S. Gabl, F. Walch, M. Sabaghian, C. Jähne, L. Hausperger, and S. Haddadin, "The franka emika robot: A reference platform for robotics research and education," *IEEE Robotics and Automation Magazine*, vol. 29, no. 2, pp. 46–64, June 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9721535

[13] F. J. Abu-Dakka and M. Saveriano, "Variable impedance control and learning—a review," *Frontiers in Robotics and AI*, vol. 7, 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frobt.2020.590681/full

[14] G. Niemeyer, C. Preusche, and G. Hirzinger, "Telerobotics," in *Springer Handbook of Robotics*, B. Siciliano and O. Khatib, Eds. Berlin, Germany: Springer, 2016, pp. 1085–1108.

[15] A. Ajoudani, N. G. Tsagarakis, and A. Bicchi, "Tele-impedance: Teleoperation with impedance regulation using a body–machine interface," *The International Journal of Robotics Research*, vol. 31, no. 13, pp. 1642–1656, 2012. [Online]. Available: https://journals.sagepub.com/doi/10.1177/0278364912464668

[16] L. Peternel, T. Petrič, and J. Babič, "Human-in-the-loop approach for teaching robot assembly tasks using impedance control interface," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 1497–1502.

[17] S. Park, W. Lee, W. K. Chung, and K. Kim, "Programming by demonstration using the teleimpedance control scheme: Verification by an semg-controlled ball-trapping robot," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 2, pp. 998–1006, Feb 2019.

[18] C. Yang, C. Zeng, P. Liang, Z. Li, R. Li, and C.-Y. Su, "Interface design of a physical human-robot interaction system for human impedance adaptive skill transfer," *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 1, pp. 329–340, Jan 2018.

[19] L. Peternel, N. Tsagarakis, and A. Ajoudani, "A human-robot comanipulation approach based on human sensorimotor information," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 7, pp. 811–822, Jul 2017.

[20] D. S. Walker, J. K. Salisbury, and G. Niemeyer, "Demonstrating the benefits of variable impedance to telerobotic task execution," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2011, pp. 1348–1353.

[21] L. Peternel, T. Petrič, and J. Babič, "Robotic assembly solution by human-in-the-loop teaching method based on real-time stiffness modulation," *Autonomous Robots*, vol. 42, no. 1, pp. 1–17, 2018.

[22] F. M. C. Kraakman, "Reduced-complefxity teleimpedance command interface enabling single-handed control of 3d stiffness for unstructured tasks," Master's thesis, Delft University of Technology, 2024. [Online]. Available: http://resolver.tudelft.nl/uuid:c69c9360-cdcd-4b61-9fc3-5e6bec67c523

[23] L. Peternel, N. Beckers, and D. A. Abbink, "Independently commanding size, shape and orientation of robot endpoint stiffness in tele-impedance by virtual ellipsoid interface," in *Proceedings of the 20th International Conference on Advanced Robotics (ICAR)*, 2021, pp. 99–106.

[24] A. Ajoudani, "Replicating human stiffness profile with a cartesian impedance controller in realtime," in *Transferring Human Impedance Regulation Skills to Robots*. Cham: Springer, 2016, pp. 33–45. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-24205-7_4

[25] Y. Michel, R. Rahal, C. Pacchierotti, P. Robuffo Giordano, and D. Lee, "Bilateral teleoperation with adaptive impedance control for contact tasks," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5429–5436, July 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9380915/

[26] Y.-C. Huang, D. A. Abbink, and L. Peternel, "A semi-autonomous teleimpedance method based on vision and voice interfaces," in *Proceedings of the 20th International Conference on Advanced Robotics (ICAR)*, 2021, pp. 180–186.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.

[28] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020. [Online]. Available: https://arxiv.org/abs/2001.08361

[29] OpenAI, "Gpt-4v(ision) system card," https://openai.com/research/gpt-4v-system-card, 09 2023, discusses the integration of vision capabilities into GPT-4, enabling the model to analyze and interpret image inputs.

[30] R. Konrad, N. Padmanaban, J. G. Buckmaster, K. C. Boyle, and G. Wetzstein, "GazeGPT: Augmenting human capabilities using gaze-contingent contextual AI for smart eyewear," *arXiv preprint arXiv:2401.17217*, 2024. [Online]. Available: https://arxiv.org/abs/2401.17217

[31] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[32] F. Liu, W. Cai, J. Huo, C. Zhang, D. Chen, and J. Zhou, "Making large vision language models to be good few-shot learners," *arXiv preprint arXiv:2408.11297*, 2024. [Online]. Available: https://arxiv.org/abs/2408.11297

[33] J. Silva-Rodríguez, S. Hajimiri, I. Ben Ayed, and J. Dolz, "A closer look at the few-shot adaptation of large vision-language models," *arXiv preprint arXiv:2312.12730*, 2023. [Online]. Available: https://arxiv.org/abs/2312.12730

[34] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a visual language model for few-shot learning," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 18 028–18 045. [Online]. Available: https://arxiv.org/abs/2204.14198

APPENDIX A
SOFTWARE ARCHITECTURE

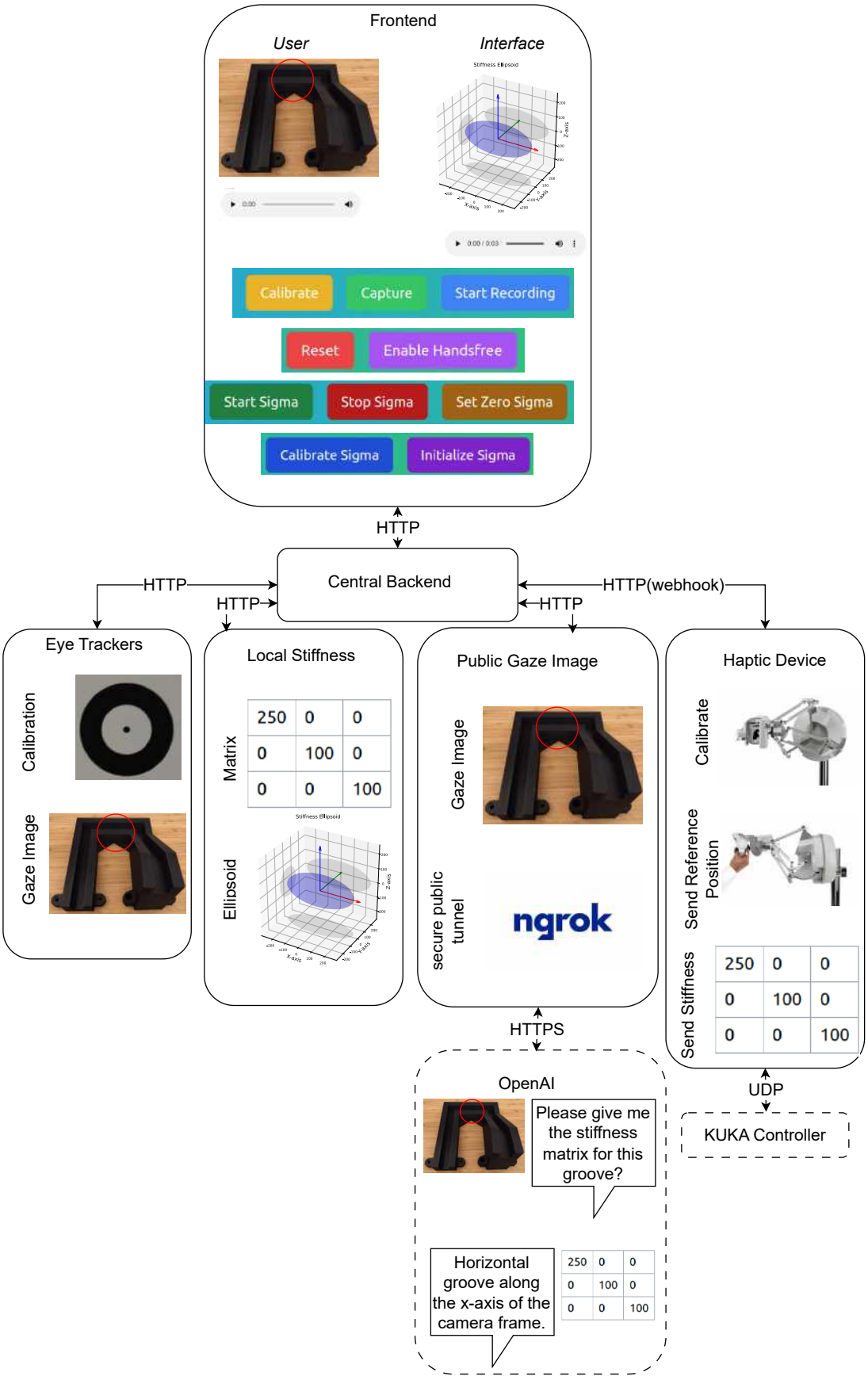

Fig. 15. Full-page software architecture diagram

APPENDIX B
DETAILED MATHEMATICAL FOUNDATIONS TELEIMPEDANCE

This appendix provides an extensive mathematical explanation of how the stiffness ellipsoid is constructed, how it influences endpoint impedance, and how it ultimately relates to the joint-level dynamics of the robot. The derivations here offer a rigorous foundation, allowing the main text to remain focused on higher-level system design and implementation details.

A. Endpoint Stiffness and Hooke's Law

In robotic interaction tasks, the relationship between the applied force and the displacement at the robot's endpoint can be modeled using an extension of Hooke's law:

$$\mathbf{F} = \mathbf{K}\mathbf{x},$$

where:

- $\mathbf{F} \in \mathbb{R}^3$ is the force vector acting on the end-effector,
- $\mathbf{K} \in \mathbb{R}^{3\times3}$ is the translational stiffness matrix,
- $\mathbf{x} \in \mathbb{R}^3$ is the displacement from the desired equilibrium position.

In our teleimpedance framework, $\mathbf{K}$ is controlled by the proposed *visio-verbal* interface, ensuring that the robot's endpoint stiffness is dynamically adjustable based on operator inputs (gaze and speech).

B. Representing $\mathbf{K}$ as a Stiffness Ellipsoid

1) Eigenvalue Decomposition of $\mathbf{K}$*:* Since $\mathbf{K}$ is assumed to be symmetric and positive semi-definite, it can be diagonalized via eigenvalue decomposition:

$$\mathbf{K} = \mathbf{Q}\mathbf{\Lambda}\,\mathbf{Q}^T,$$

where:

- $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3)$ contains the eigenvalues of $\mathbf{K}$.
- $\mathbf{Q} \in \mathbb{R}^{3\times3}$ is an orthogonal matrix whose columns are the normalized eigenvectors $\{\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3\}$.

Each eigenvalue λ_i is non-negative (since $\mathbf{K}$ is positive semi-definite) and corresponds to a principal stiffness direction. The eigenvectors $\mathbf{q}_i$ give the orientation of these principal axes in the robot's Cartesian space.

2) Ellipsoid Definition: A stiffness ellipsoid is defined implicitly by:

$$\big\{\mathbf{x} \in \mathbb{R}^3 \mid \mathbf{x}^T\mathbf{K}\,\mathbf{x} = 1\big\}.$$

Substituting $\mathbf{K} = \mathbf{Q}\mathbf{\Lambda}\,\mathbf{Q}^T$ yields:

$$\mathbf{x}^T\mathbf{Q}\mathbf{\Lambda}\,\mathbf{Q}^T\mathbf{x} = 1 \quad\Longleftrightarrow\quad \sum_{i=1}^{3} \lambda_i\,(\mathbf{q}_i^T\mathbf{x})^2 = 1.$$

Hence, the ellipsoid's principal axes align with the eigenvectors $\mathbf{q}_i$, and the axis lengths are inversely proportional to $\sqrt{\lambda_i}$. Thus, larger λ_i indicates stiffer behavior along that axis, yielding a shorter ellipsoid radius in that direction.

C. Connection to the Robot's Impedance Controller

In Cartesian space, the robot's impedance controller regulates the endpoint force $\mathbf{F}$ using:

$$\mathbf{F} = \mathbf{K}_{\text{robot}}(\mathbf{x}_d - \mathbf{x}_a) + \mathbf{D}\big(\dot{\mathbf{x}}_d - \dot{\mathbf{x}}_a\big),$$

where:

- $\mathbf{x}_a, \dot{\mathbf{x}}_a$ denote the actual pose and velocity of the end-effector,
- $\mathbf{x}_d, \dot{\mathbf{x}}_d$ denote the desired (reference) pose and velocity,
- $\mathbf{K}_{\text{robot}} \in \mathbb{R}^{6\times6}$ is the complete stiffness matrix (including both translational and rotational components),
- $\mathbf{D} \in \mathbb{R}^6$ (or $\mathbb{R}^{6\times6}$, depending on notation) represents the damping.

In our design, we assume $\mathbf{K}_{\text{rotation}}$ (the rotational component) is fixed, leaving $\mathbf{K}_{\text{translational}} \in \mathbb{R}^{3\times3}$ subject to real-time updates from the *visio-verbal* interface. This approach suits tasks like slide-in-the-groove, where adapting translational stiffness alone proves sufficient for improved performance.

D. Joint-Level Dynamics and Robot Control

While the impedance controller operates in Cartesian space, the underlying robot dynamics are governed in joint space. Let $\mathbf{q} \in \mathbb{R}^n$ denote the joint angles (with n typically 6 or 7 for a manipulator), then the robot's equation of motion can be written as:

$$\mathbf{M}(\mathbf{q})\,\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})\,\dot{\mathbf{q}} + \mathbf{g}(\mathbf{q}) + \mathbf{J}^T(\mathbf{q})\,\mathbf{F} = \boldsymbol{\tau},$$

where:

- $\mathbf{M}(\mathbf{q})$ is the joint-space mass (inertia) matrix,
- $\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})$ captures Coriolis and centrifugal effects,
- $\mathbf{g}(\mathbf{q})$ is the gravity torque,
- $\mathbf{J}(\mathbf{q})$ is the Jacobian mapping joint velocities to Cartesian velocities,
- $\mathbf{F} \in \mathbb{R}^6$ includes the interaction force/torque at the end-effector,
- $\boldsymbol{\tau} \in \mathbb{R}^n$ is the vector of joint torques.

Here, $\mathbf{F}$ comprises both the task-related forces $\mathbf{F}_{\text{task}}$ and the impedance forces $\mathbf{F}_{\text{imp}}$, ensuring that the end-effector behavior aligns with the stiffness parameters set by the visio-verbal interface.

E. Concluding Remarks

By unifying the stiffness ellipsoid representation in Cartesian space with joint-space dynamics, this framework offers a powerful and intuitive approach to teleimpedance. The *visio-verbal* interface dynamically updates $\mathbf{K}_{\text{translational}}$, ensuring the robot's endpoint compliance is continually optimized for the given task—*e.g.*, slide-in-the-groove—without burdening the operator with complex calibration or additional hardware. Eigenvalue decomposition simplifies stiffness visualization, while the underlying Cartesian impedance controller and joint-level equations ensure stable, precise motion execution.

Note: This appendix summarizes the mathematical underpinnings. See Section II for a practical discussion on visualizing stiffness ellipsoids within our teleimpedance framework.

APPENDIX C
STIFFNESS MATRIX DECOMPOSITIONS FOR THE GROOVE STRUCTURE

This appendix provides detailed eigenvalue decompositions for the stiffness matrices used in different sections of the groove structure (Figure 7 in the main text). Each decomposition illustrates how the matrix's principal axes (eigenvectors) and principal stiffnesses (eigenvalues) define the shape and orientation of the corresponding stiffness ellipsoid.

A. Entrance

The stiffness matrix for the entrance is the diagonal matrix:

$$\mathbf{K}_{\text{entrance}} = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 250 \end{bmatrix}.$$

1) Eigenvalue Decomposition: Since $\mathbf{K}_{\text{entrance}}$ is already diagonal:

$$\mathbf{\Lambda} = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 250 \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The principal axes align with the coordinate axes, but the z-axis is stiffer at 250 N/m, leading to an "elongated" shape along z.

2) Ellipsoid Plot: At the entrance, the stiffness ellipsoid is elongated along z as presented in figure 16:

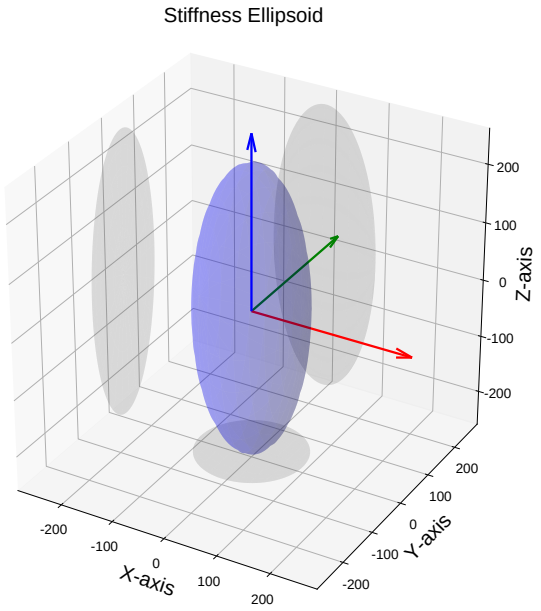

Fig. 16. Stiffness ellipsoid for the entrance, with $\mathbf{K} = \text{diag}(100, 100, 250)$.

B. y path

The groove section right after the entrance is oriented along the y-axis, and therefore requires higher stiffnes in that direction with stiffness matrix matrix:

$$\mathbf{K}_{y\text{-path}} = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 250 & 0 \\ 0 & 0 & 100 \end{bmatrix}.$$

1) Eigenvalue Decomposition: Again, $\mathbf{K}_{y\text{-path}}$ is diagonal:

$$\mathbf{\Lambda} = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 250 & 0 \\ 0 & 0 & 100 \end{bmatrix}, \quad \mathbf{Q} = \mathbf{I}_{3\times 3}.$$

The principal axes align with the coordinate axes, but the y-axis is stiffer at 250, leading to an "elongated" shape along y.

2) Ellipsoid Plot: Figure 17 shows the ellipsoid, which is elongated in the y direction compared to the x and z directions.

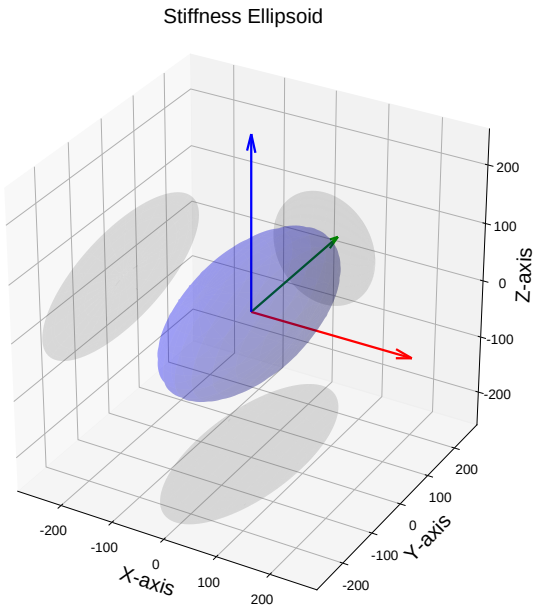

Fig. 17. Stiffness ellipsoid for the y-path, $\text{diag}(100, 250, 100)$.

C. x-path

This matrix is a 90° rotated version of the y-path:

$$\mathbf{K}_{x\text{-path}} = \begin{bmatrix} 250 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 100 \end{bmatrix}.$$

Again, diagonal but now with a higher eigenvalue of 250 N/m on the x axis.

1) Eigenvalue Decomposition:

$$\mathbf{\Lambda} = \begin{bmatrix} 250 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 100 \end{bmatrix}, \quad \mathbf{Q} = \mathbf{I}_{3\times 3}.$$

Higher stiffness occurs along the x axis.

2) Ellipsoid Plot: The ellipsoid for x-path elongates along x relative to the y and z axes. Figure 18 illustrates its shape.

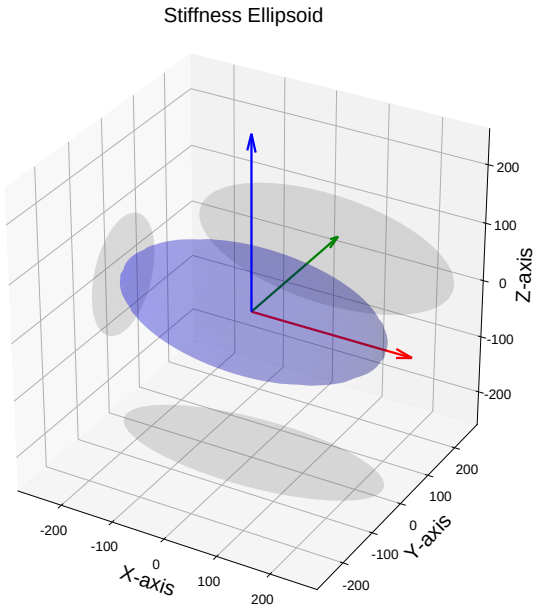

Fig. 18. Stiffness ellipsoid for the x-traverse, $\mathrm{diag}(250,\ 100,\ 100)$.

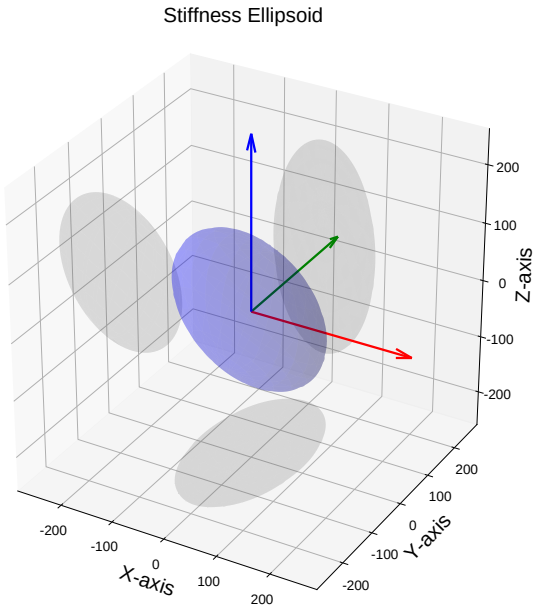

Fig. 19. Stiffness ellipsoid for the slant region, illustrating a rotation in the y-z plane.

D. Slant

The "slant" section is more complex due to off-diagonal terms:

$$\mathbf{K}_{\text{slant}} = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 175 & -75 \\ 0 & -75 & 175 \end{bmatrix}.$$

Here, x remains decoupled with 100 N/m , but y and z are coupled via off-diagonal elements of -75 N/m.

1) Eigenvalue Decomposition: Let

$$\mathbf{K}_{\text{slant}} = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 175 & -75 \\ 0 & -75 & 175 \end{bmatrix}.$$

It can be shown that the eigenvalues are $[100,\ 100,\ 250]$, and the eigenvectors are rotated relative to the y-z axes with 45 degrees.

Denoting $\mathbf{\Lambda}_{\text{slant}} = \mathrm{diag}(100,\ 100,\ 250)$ and $\mathbf{Q}_{\text{slant}} \in \mathbb{R}^{3\times3}$ as the matrix of orthonormal eigenvectors, we have:

$$\mathbf{K}_{\text{slant}} = \mathbf{Q}_{\text{slant}}\, \mathbf{\Lambda}_{\text{slant}}\, \mathbf{Q}_{\text{slant}}^{T}.$$

The eigenvectors reflect a rotation in the y-z plane due to the off-diagonal -75 terms.

2) Ellipsoid Plot: Figure 19depicts the ellipsoid after applying the rotation $\mathbf{Q}_{\text{slant}}$. Notably, two axes share the same stiffness (100) but are rotated relative to the global axes, while one axis has a higher stiffness (250).

E. Summary of Matrix Groupings

Table II summarizes each matrix and its key eigenvalue characteristics:

TABLE II
SUMMARY OF GROOVE STIFFNESS MATRICES AND PRINCIPAL EIGENVALUES

Location	Stiffness Matrix	Eigenvalues	Diagonal/Off-Diagonal?
Entrance	$\mathrm{diag}(100, 100, 250)$	$(100, 100, 250)$	Diagonal
Y-path	$\mathrm{diag}(100, 250, 100)$	$(100, 250, 100)$	Diagonal
X-path	$\mathrm{diag}(250, 100, 100)$	$(250, 100, 100)$	Diagonal
Slant	$\begin{bmatrix} 100 & 0 & 0 \\ 0 & 175 & -75 \\ 0 & -75 & 175 \end{bmatrix}$	$(100, 250, 100)$	Off-Diagonal in y-z

This grouping reflects that the system has distinct "modes" of stiffness (elongated in x, y, or z or along a direction 45 degrees rotated within the y-z plane), each corresponding to a specific region of the groove structure. The visio-verbal impedance interface allows the operator to seamlessly switch between these configuration using voice and gaze.

In this appendix, we have carried out explicit eigenvalue decompositions for each stiffness matrix used within the groove structure. By plotting the corresponding ellipsoids during teleoperation, operators are able to quickly check the current stiffness configuration using these visualizations. This appendix also provides the reader with this visual understanding of the stiffness matrices for the sections in the groove structure.

APPENDIX D
PROMPT OPTIMIZATION - EXPERIMENT DETAILS

This appendix gives additional results from the prompt engineering experiment.

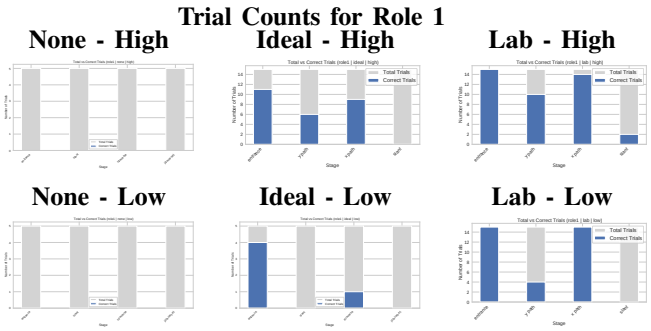

Fig. 20. Trial count distribution for Role 1 under different prior conditions and resolution settings.

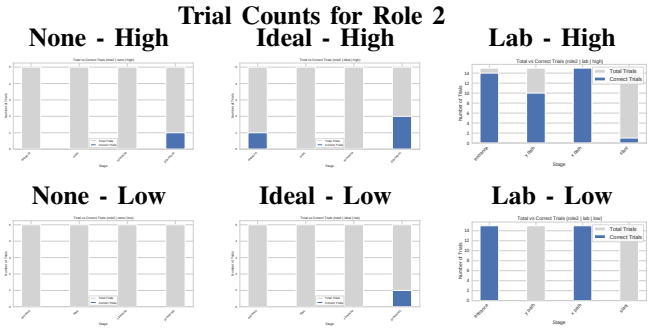

Fig. 21. Trial count distribution for Role 2 under different prior conditions and resolution settings.

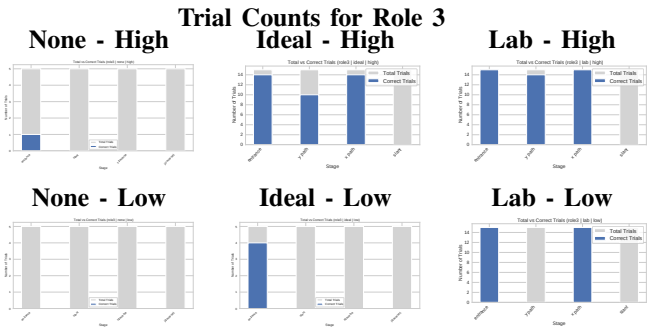

Fig. 22. Trial count distribution for Role 3 under different prior conditions and resolution settings.

APPENDIX E
PRIOR MESSAGE LIST IMAGES

Figure 23 presents the ideal and lab setup data and the test data that were used to aquire the results based on the first experiment described in section III.

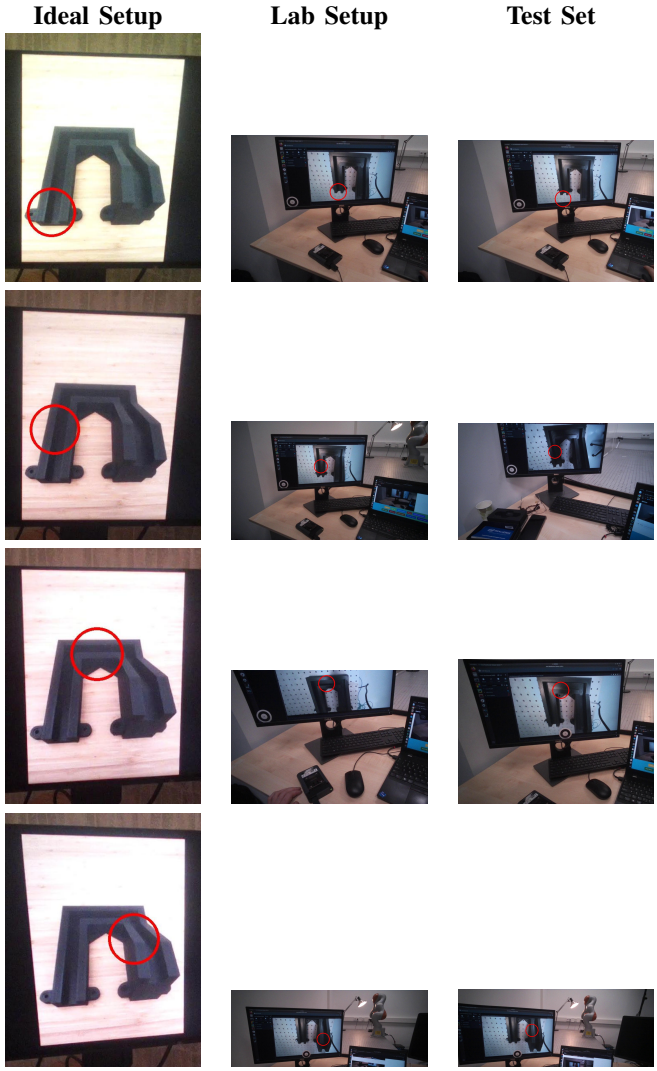

Fig. 23. Listing of images used in different experimental conditions. The first column represents images captured in the ideal setup for few shot prompting, the second column represents images from the lab setup used for few shot prompting, and the third column contains the test set images used for evaluation of the different prompt combinations.

APPENDIX F
TASK DESCRIPTIONS

This appendix presents the task descriptions (also knows as system roles in the deep learning field) used in the experiment. Each role builds upon the previous one, progressively adding more context and structured guidance to the vision-language model (VLM) for generating the stiffness matrix.

A. *Role 1*

System Role Content: You are able to analyze and process images. You are an interface designed to determine the stiffness matrix. Unless instructed otherwise, the only output you should provide is a 3 by 3 stiffness matrix formatted using the structure below **without any extra text or comments** between the header and the matrix.

Stiffness Matrix Format:

$$\mathbf{K} = \begin{bmatrix} K_{xx} & K_{yx} & K_{zx} \\ K_{xy} & K_{yy} & K_{zy} \\ K_{xz} & K_{yz} & K_{zz} \end{bmatrix}$$

You must determine this stiffness matrix based on the most recently provided image. The conversation history contains prior messages simulating an interaction between you and the user, each associated with an image and its corresponding stiffness matrix. If such references exist, use them to derive the stiffness matrix for the current image.

B. Role 2

System Role Content: You are able to analyze and process images. You are an interface designed to determine the stiffness matrix. The only output you should provide is a 3 by 3 stiffness matrix formatted using the structure:

$$\mathbf{K} = \begin{bmatrix} K_{xx} & K_{yx} & K_{zx} \\ K_{xy} & K_{yy} & K_{zy} \\ K_{xz} & K_{yz} & K_{zz} \end{bmatrix}$$

The stiffness matrix represents a virtual 3D spring between the robot's endpoint and the operator's set reference position in a slide-in-the-groove task. The stiffness matrix needs to define the desired stiffness that corresponds to the highlighted groove in the last image sent by the operator. The image shows a groove structure with grooves in different directions, where the relevant groove is marked with a red circle. Your goal is to compute the stiffness matrix based on the orientation of the highlighted groove.

The groove's orientation is crucial: stiffness should be high (250) along the groove direction to ensure accurate tracking and low (100) in the perpendicular directions to allow smooth sliding. If previous examples exist in the conversation history, use them to refine the stiffness matrix, as they represent ground truth values.

C. Role 3

System Role Content: You are able to analyze and process images. You are an interface designed to determine the stiffness matrix. The only output you should provide is a 3 by 3 stiffness matrix formatted as:

$$\mathbf{K} = \begin{bmatrix} K_{xx} & K_{yx} & K_{zx} \\ K_{xy} & K_{yy} & K_{zy} \\ K_{xz} & K_{yz} & K_{zz} \end{bmatrix}$$

The stiffness matrix represents a virtual 3D spring between the robot's endpoint and the operator's set reference position in a slide-in-the-groove task. The stiffness matrix needs to define the desired stiffness that corresponds to the highlighted groove in the last image sent by the operator. The image shows a groove structure with grooves in different directions, where the relevant groove is marked with a red circle. Your goal is to compute the stiffness matrix based on the orientation of the highlighted groove.

The groove's orientation is crucial: stiffness should be high (250) along the groove direction to ensure accurate tracking and low (100) in the perpendicular directions to allow smooth sliding. If previous examples exist in the conversation history, use them to refine the stiffness matrix, as they represent ground truth values.

Predefined Stiffness Matrices for Groove Orientations:

Groove along X-axis (left-to-right):

$$\mathbf{K} = \begin{bmatrix} 250 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 100 \end{bmatrix}$$

Groove along Y-axis (bottom-to-top):

$$\mathbf{K} = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 250 & 0 \\ 0 & 0 & 100 \end{bmatrix}$$

Groove diagonal in the YZ-plane:

$$\mathbf{K} = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 175 & -75 \\ 0 & -75 & 175 \end{bmatrix}$$

Entrance at the left-bottom of the structure:

$$\mathbf{K} = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 250 \end{bmatrix}$$

If the groove does not match any of these predefined cases, determine the stiffness matrix by analyzing the groove's orientation and applying the appropriate high (250) and low (100) stiffness values accordingly.

APPENDIX G
HARDWARE COMPONENTS

TABLE III
OVERVIEW OF HARDWARE COMPONENTS

Component	Details	Notes
Mobile Eye-Trackers	Tobii Pro Glasses 2	• **Reason for Choice:** Eye tracking capability; available at the lab. A static display-based eye tracker would have been preferred for better stability. • **Connection Details:** Communicates with the laptop via an external USB Wi-Fi adapter, as the internal Wi-Fi adapter is reserved for internet access to send requests to OpenAI.
External Wi-Fi Adapter	TP-Link USB Wi-Fi Adapter	• **Reason for Choice:** Needed to connect the eye tracker to the laptop while freeing the internal Wi-Fi adapter for internet access. • **Connection Details:** Connected to the laptop via USB and to the eyetrackers via WIFI.
Sigma7 Device	Force Dimension sigma7	• **Reason for Choice:** Available at the TU Delft robotics lab; provides high-precision haptic feedback for teleoperation tasks. • **Connection Details:** Connected to the laptop via USB.
KUKA Robot Arm	KUKA LBR iiwa	• **Reason for Choice:** Available at the lab; collaborative design with 7 DoF. • **Connection Details:** Connected to a switch via Ethernet, which links the robot arm to a desktop running the KUKA controller.
Desktop Hosting KUKA Controller	Dell Precision Workstation	• **Reason for Choice:** Required to interface with the KUKA robot arm. • **Connection Details:** Connected to the switch via Ethernet. Displayed on the monitor via HDMI.
Camera	Depth Camera (e.g., Intel RealSense)	• **Reason for Choice:** Already available in the lab with a 3D-printed mount for the robot. • **Connection Details:** Mounted on the robot's endpoint and connected to the desktop hosting the KUKA controller via USB.
Display Screen	Dell Monitor	• **Reason for Choice:** Already positioned in the lab. • **Connection Details:** Connected to the desktop hosting the KUKA controller via HDMI.
Switch	D-Link Ethernet Switch	• **Reason for Choice:** Available in the lab for networking multiple devices. • **Connection Details:** Connects the laptop, robot arm, and desktop hosting the KUKA controller via Ethernet.
Computational Setup	Laptop (Lenovo P15v)	• **Reason for Choice:** Personal laptop; communicates with OpenAI over the internet via the internal Wi-Fi adapter. • **Connection Details:** Connected to a switch via Ethernet to communicate with the KUKA controller, connected to the external wifi adapter via USB and connected to the Sigma7 via USB.