Unimanual vs. Bimanual Gesture Control for Pick-and-Place in Simulation

Maria BOUSSENAH, Wojciech ŁOBODA

January 15, 2026

Abstract

We study whether bimanual gesture input improves performance and robustness over unimanual input for a robotic pick-and-place task in simulation. Using vision-based hand tracking and a four-region screen mapping, we compare unimanual control, where a single hand sequentially controls all joints, and bimanual control, where the left and right hands control disjoint joint subsets in parallel. We evaluate task time, placement accuracy, and success-related metrics under varying task difficulty. Completion times were similar across modalities for easy and medium trials, with bimanual control trending faster for hard trials. Failure rates were comparable, while bimanual control achieved lower placement error and higher subjective confidence.

1 Introduction

Robotic manipulation is increasingly used in domains where flexibility and human oversight are required, such as remote operation, training, and assistive robotics. In these contexts, the way humans interact with robots plays a crucial role in overall system effectiveness [2]. Traditional teleoperation interfaces often rely on joysticks, keyboards, or specialized devices, which can impose a steep learning curve and limit the naturalness of interaction.

Gesture-based teleoperation has emerged as a promising alternative, allowing users to control robotic systems through natural hand movements. In particular, vision-based hand tracking enables robot control using commodity RGB cameras, reducing hardware requirements while supporting intuitive interaction [5, 6]. Recent advances in computer vision have made real-time hand pose estimation sufficiently robust for interactive human–robot systems, making gesture-driven robot control increasingly accessible.

However, controlling a multi-degree-of-freedom robotic arm using hand gestures remains challenging. A central design question concerns how control should be distributed across the user's hands. Many gesture-based systems rely on unimanual input, requiring the user to sequentially switch between different control dimensions. While simple, this approach may limit efficiency and precision for complex manipulation tasks [2]. Alternatively, bimanual input allows control to be distributed across two hands, potentially enabling parallel action and improved coordination, but at the cost of increased cognitive and motor demands [3].

Despite the intuitive appeal of bimanual interaction, its benefits for gesture-based robot teleoperation are not yet well understood, particularly under controlled conditions where interface layout and task demands are held constant. This motivates a systematic comparison of unimanual and bimanual gesture control for robotic manipulation.

In this work, we present an empirical study comparing unimanual and bimanual vision-based gesture control for a pick-and-place task performed by a single robotic arm in simulation. Using the same hand-tracking pipeline, identical four-region screen mapping, and matched task conditions, we evaluate how distributing control across one versus two hands affects task time, placement accuracy, and robustness under varying levels of task difficulty.

2 Related Work

Vision-based hand gesture control for robots has received considerable attention because it enables intuitive interaction without specialized hardware. Recent survey work reviews hand detection, tracking, and gesture recognition techniques for human–robot interaction, highlighting the role of RGB and RGB-D cameras in translating natural hand motions into robot commands [5]. These methods form the perception foundation for vision-based teleoperation systems.

Several practical systems have demonstrated the feasibility of gesture-based robot control using commodity cameras. For example, Wang et al. proposed a MediaPipe-based hand gesture tracking framework for teleoperating a mobile manipulator in simulation, showing that hand position and discrete gestures can be effectively mapped to robot motion commands in PyBullet [8]. Other vision-based teleoperation approaches similarly show that users can perform pick-and-place and navigation tasks without wearable sensors, emphasizing the accessibility of camera-only solutions [1, 7].

More recently, research has explored more complex teleoperation architectures that support multi-degree-of-freedom control and coordinated manipulation. Zhu et al. introduced a bimanual teleoperation architecture with dual robotic arms and anthropomorphic grippers, demonstrating that distributing control across two hands can improve manipulation capability in unstructured environments [9]. Related work also explores hybrid gesture and upper-limb pose mappings

for dual-arm teleoperation, enabling more flexible operator intervention during coordinated actions [4].

However, such systems often rely on specialized hardware or focus on dual-arm robots, making it difficult to isolate the effect of unimanual versus bimanual input mappings under controlled conditions. Moreover, few studies directly compare unimanual and bimanual vision-based gesture control using the same visual interface and task setup. Our work addresses this gap by systematically comparing unimanual and bimanual gesture control for a simulated pick-and-place task under an identical four-region interface and commodity webcam tracking.

3 System Description

3.1 Task and Environment

We implement a pick-and-place task in PyBullet using a Franka Panda robotic arm. In each trial, the participant must (1) reach and grasp a cube using a pinch gesture, (2) transport the cube to a target marker, and (3) release the cube near the target to complete the trial.

Figure 1 shows how the simulation environment looks like. It consists of two tables (pick and place), a cube, a cylindrical target marker, and visual aids including a halo around the cube and projected shadows to aid depth perception.

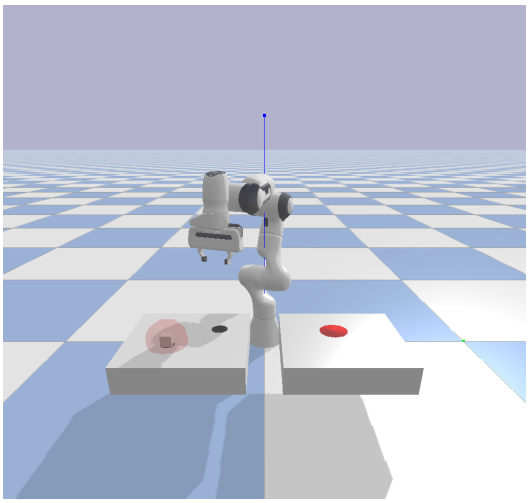

Figure 1: Simulation Environment.

3.2 Gesture Sensing and Mapping

Hand landmarks are tracked in real time using MediaPipe Hands. The camera frame is divided into four vertical regions corresponding to four robot joints:

$$\text{BASE} \mid \text{SHOULDER} \mid \text{WRIST} \mid \text{ELBOW}.$$

The camera feed displays annotated regions, landmarks, and additional user interface elements that are visible to the participant and assist with navigation, as illustrated in Figure 2.

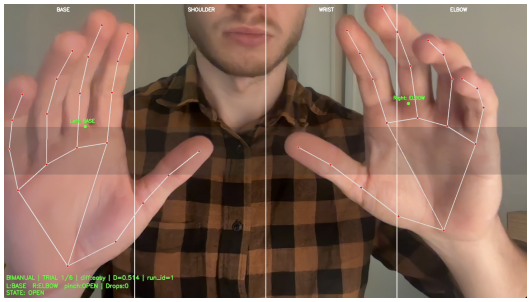

Figure 2: The Camera Feed and User Interface.

Vertical hand motion outside a central deadzone generates joint velocity commands. Upward motion produces positive velocity and downward motion produces negative velocity. Grasping is controlled using a pinch gesture (thumb–index distance) with hysteresis and a delayed release to reduce accidental drops.

Unimanual control: A single hand controls all four joints. The horizontal position of the hand selects the active joint region, and vertical motion controls joint velocity. Only one joint is actively controlled at a time.

Bimanual control: Two hands can be tracked simultaneously. The left hand controls the **BASE** and **SHOULDER** joints (left two screen regions), while the right hand controls the **WRIST** and **ELBOW** joints (right two regions). Both hands may issue commands concurrently, enabling parallel joint control. A short command-hold mechanism is used to reduce flicker when hands momentarily leave the active regions.

3.3 Assisted Grasp in Simulation

To reduce sensitivity to physics instabilities and isolate interaction effects, we employ an assisted grasp mechanism. When the gripper closes while the end-effector is within fixed XY and Z thresholds of the cube, a fixed constraint is created that preserves the cube's relative pose to the end-effector. This prevents snapping artifacts and unintended slips during transport. The constraint is removed upon release.

4 Experimental Design

4.1 Conditions and Trial Structure

We compare two control modes: **Unimanual** and **Bimanual**. Each participant completes both modes in a counterbalanced order to mitigate learning effects.

Task difficulty is manipulated via cube placement difficulty, with three levels:

- **Easy**: cube positioned centrally and close to the robot.

- **Medium**: cube at a nominal baseline position.
- **Hard**: cube offset laterally and further from the neutral reach.

Each difficulty level is repeated twice per mode, resulting in:

$$3 \text{ difficulties} \times 2 \text{ repetitions} = 6 \text{ trials per mode.}$$

Trial order is randomly shuffled and shared between unimanual and bimanual blocks for each participant.

4.2 Participants

Ten university students participated in the experiment, all right-handed. All participants reported prior exposure to hand-recognition systems through consumer technologies such as gesture-controlled user interfaces, camera-based gaming systems, and virtual or augmented reality applications. None had prior experience with vision-based robot teleoperation.

4.3 Counterbalancing

Participants are automatically assigned to one of two groups based on participant index parity. Group A performs the unimanual condition first, followed by the bimanual condition, while Group B performs the conditions in reverse order. Participant identifiers and condition order are generated automatically and persist across program restarts.

4.4 Metrics

The following objective metrics are logged per trial:

- **Time Total**: time from trial start to cube release.
- **Grasp Time**: time from trial start to first successful grasp.
- **Placement Error**: 2D distance between cube and target at release.
- **Success**: whether placement error is below a fixed threshold.
- **Difficulty**: cube placement difficulty level.

4.5 Questionnaire

After completing the trials, participants are asked to complete a questionnaire for each input modality. The questionnaire consists of statements evaluated using a five-point Likert scale, as well as self-assessment question.

The following were included:

- How would you rate your performance? (1 = very good, 5 = very poor)
- The system is easy to learn (1 = strongly disagree, 5 = strongly agree).
- The system is hard to use (1 = strongly disagree, 5 = strongly agree).
- I felt confident using the system (1 = strongly disagree, 5 = strongly agree) .

5 Results

5.1 Quantitative Results

Data was collected during the sessions in which 10 participants tested the system. Both metrics and subjective questionnaire responses were recorded.

Figure 3 shows the mean total times and grasp times for the succesful trials. Overall, the patterns are similar for the two metrics. Slightly higher total times were observed for the bimanual modality on easy and medium difficulty trials, whereas for hard difficulty, the bimanual modality appears slightly faster. The results between easy and medium difficulties are similar. The confidence intervals indicate strong variability in the data.

Mean percentage of failed trials per participant for both modalities was presented in Figure 4. For both modalities around 30% of attempted trials were failed by average. The percentages are similar across modalities, and the wide confidence intervals suggest limited precision in the estimates.

Figure 5 shows the mean placement error for both modalities. The error is lower for the bimanual modality compared to the unimanual modality. The confidence intervals indicate strong variability as for previous results, especially for the unimanual modality.

The results were consistent between the groups that started with unimodal and bimanual trials.

5.2 Qualitative Results

The answers to the questionaire are presented in Figure 6. The participants rated their performance higher when using the bimanual modality compared to the unimanual modality. Responses regarding how easy the system was to learn were similarly distributed for both modalities, with most ratings at 3 or above.

Participants reported the unimanual modality as harder to use, with most responses being 4 or 5. For the bimanual modality, most participants rated the system as easier to use, with most responses being 2.

Regarding confidence, most participants reported feeling confident when using the bimanual modality, whereas confidence levels were lower for the unimanual modality.

Overall, the questionnaire responses suggest that participants perceived better performance, greater ease of use, and higher confidence when interacting with the bimanual system, while both modalities were generally rated as not difficult to learn.

6 Discussion and Conclusion

6.1 Interpretation

This study compared unimanual and bimanual vision-based gesture control for a simulated pick-and-place task using the same four-region interface and tracking pipeline. Overall,

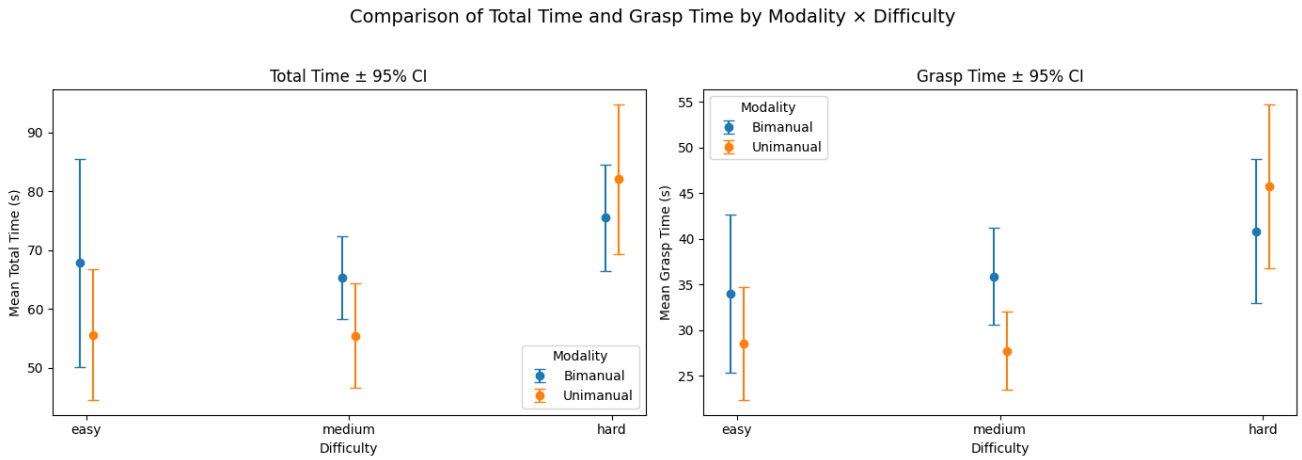

Figure 3: Grasp Times for Successful Trials per Modality and Difficulty.

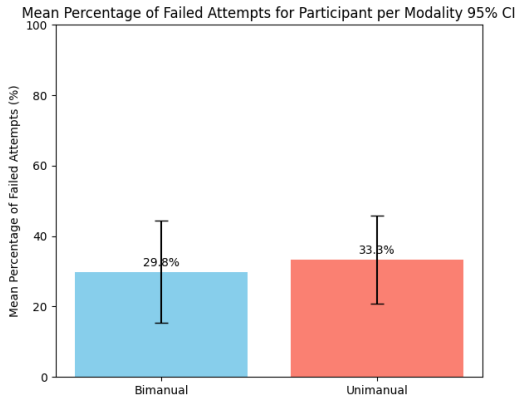

Figure 4: Mean Percentage of Failed Attempts for Participant per Modality.

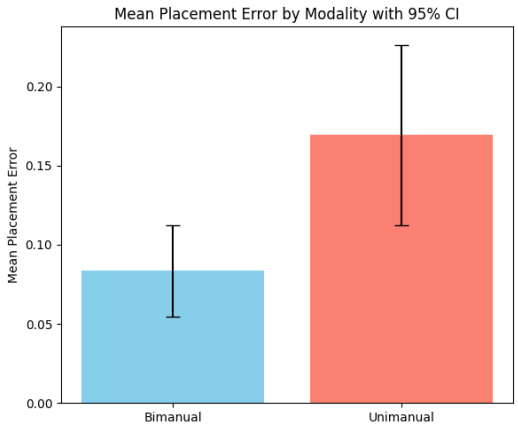

Figure 5: Mean Placement Error per Modality.

the quantitative results show a trade-off: bimanual input did not consistently improve speed, but it tended to improve precision and was perceived more positively by participants.

In terms of efficiency, mean completion times (and grasp times for successful trials) were similar across modalities for easy and medium trials, with bimanual control sometimes slightly slower. For the hard condition, bimanual control showed a modest tendency toward faster completion. This suggests that parallel control may become helpful when the task requires larger or more sustained coordination, while for simpler trials the effort of coordinating both hands may reduce any time advantage.

Robustness outcomes were comparable. The proportion of failed trials remained around 30% for both modalities, and variability across participants was high. In contrast, placement accuracy differed more clearly: bimanual control achieved lower placement error on average. This indicates that even when failures still occurred (e.g., missed placements), bimanual input supported more accurate final positioning once the cube was transported.

The questionnaire responses align with the accuracy trend. Participants reported higher confidence and perceived performance with bimanual control, and rated unimanual control as harder to use. Together, the findings suggest that bimanual gesture control may be most valuable for improving precision and perceived controllability, even when speed improvements are limited and depend on task difficulty.

6.2 Limitations

Several limitations should be considered when interpreting these results. First, the experiment was conducted in simulation with an assisted grasp mechanism, which simplifies contact dynamics and may not fully reflect real-world manipulation challenges. Second, the sample size was relatively small, leading to large confidence intervals and limiting statistical power. Third, vision-based hand tracking is sensitive to lighting conditions and hand visibility, which may have introduced variability across trials. Finally, although counterbalancing was applied, learning effects and fatigue cannot

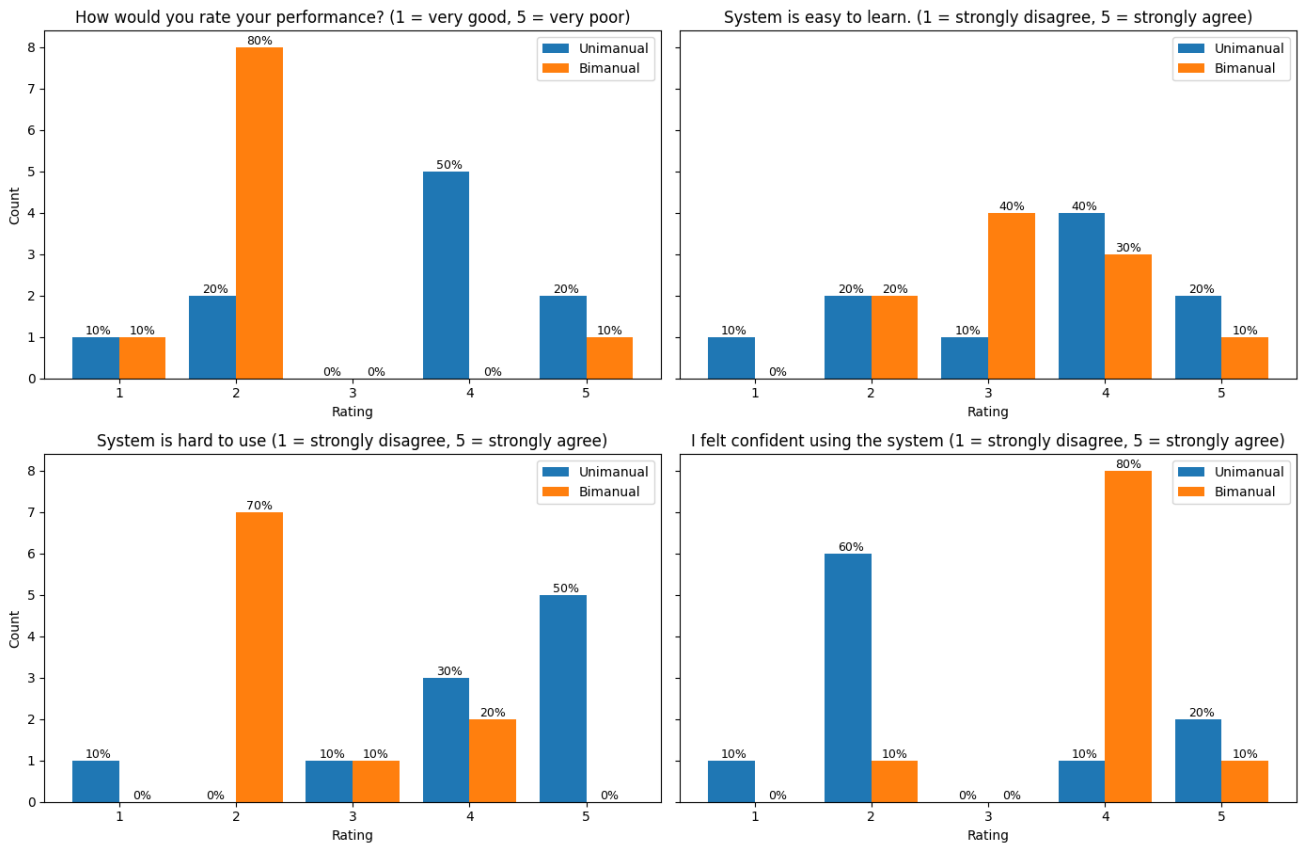

Figure 6: Responses for the questionnaire.

be entirely ruled out.

6.3 Design Recommendations

Based on limitations observed in our current implementation, we outline several recommendations for improving future vision-based gesture teleoperation systems:

- **Improve robustness to tracking variability:** Hand tracking sometimes failed when lighting changed or when hands were partially occluded. Future systems could reduce sudden robot movements when tracking becomes unreliable, for example by pausing motion until the hand is clearly detected again.
- **Adapt control to task phases:** Since bimanual control mainly improved precision rather than speed, future interfaces could switch between unimanual control for coarse motion and bimanual control for fine positioning to reduce cognitive load.
- **Enhance visual feedback for joint engagement:** While the four-region layout provides a simple control scheme, users must continuously infer which joints are active. Explicit visual cues (highlighted regions/joint indicators on the robot) could improve transparency and reduce control uncertainty.
- **Support variable motion speed:** Allowing users to modulate joint speed, particularly during transport after grasping, could improve efficiency without compromising placement accuracy.
- **Extend evaluation beyond simulation:** Testing the control mappings on a physical robot would help assess robustness under real-world contact dynamics and sensing noise.

6.4 Conclusion

This work presented a controlled comparison of unimanual and bimanual gesture-based robot teleoperation for a simulated pick-and-place task. While bimanual control did not consistently reduce task completion time, it improved placement accuracy and was perceived by users as easier and more confidence-inspiring, particularly for harder tasks. These results suggest that bimanual gesture input can enhance the quality of interaction even when objective speed gains are limited. Future work will explore more complex tasks, real robot deployment, and adaptive or personalized gesture mappings to further understand when and how bimanual control provides the greatest benefit.

References

[1] Gerardo García-Gil, Gabriela del Carmen López-Armas, and Jr. José de Jesús Navarro. Human-machine interaction: A vision-based approach for controlling a robotic hand through human hand movements. *Technologies*, 2025.

[2] Michael A. Goodrich and Alan C. Schultz. Human-robot interaction: a survey. *Foundations and Trends in Human-Computer Interaction*, 1(3):203–275, 2008.

[3] Yves Guiard. Asymmetric division of labor in human skilled bimanual action: The kinematic chain as a model. *Journal of Motor Behavior*, 19(4):486–517, 1987.

[4] Geng Yang Huayong Yang Honghao Lyu. Bimanual human-motion based robot teleoperation. In *Human Motion Awareness and Robot Teleoperation*. 2025.

[5] Jing Qi, Li Ma, Zhenchao Cui, and Yushu Yu. Computer vision-based hand gesture recognition for human-robot interaction: a review. *Complex & Intelligent Systems*, 2023.

[6] S. Rajaraman et al. A survey on vision-based hand tracking and gesture recognition for human-computer interaction. *ACM Computing Surveys*, 2021. Use as a general placeholder citation for modern hand tracking; replace with the exact paper you cite.

[7] Tianyu Wang, Yuhui Wan, Christopher Peers, Jingcheng Sun, and Chengxu Zhou. Vision-based gesture tracking for teleoperating mobile manipulators. *UK Robotics and Autonomous Systems Conference Proceedings*, 2022.

[8] Tianyu Wang, Yuhui Wan, Christopher Peers, Jingcheng Sun, and Chengxu Zhou. Vision-based gesture tracking for teleoperating mobile manipulators. In *Proceedings of the UK Robotics and Autonomous Systems Conference (UKRAS)*, pages 52–53, 2022.

[9] Guoniu Zhu, Yifan Zhang, Xiaoyu Liu, and Hongbo Wang. A bimanual robotic teleoperation architecture with anthropomorphic hybrid grippers. *Applied Sciences*, 14(3), 2024.

Appendix

A Initial Approaches for the Study

Before the final system used for the study was selected, several earlier iterations of the system were developed. These iterations are presented in this section.

A.1 3D-printed Robot Arm

Initially, instead of using a simulation, we tried to conduct the experiment using a 3D-printed robot arm, see Figure 7. After the system was prepared and the robot arm was programmed, several issues arose during the testing phase that made it impossible to use the physical setup for the final experiments.

The robot arm required additional programming and adjustments for our experiment. However, multiple hardware problems occurred, including parts breaking down and requiring repair, which made the process too time-consuming. Conducting the experiment with a physical robot arm also posed logistic challenge, as participants would have needed to perform the tasks in a controlled environment, specific room for example.

With these issues in mind, we decided to use a simulation environment instead. This approach allowed us to focus on the actual experiment without being constrained by hardware limitations.

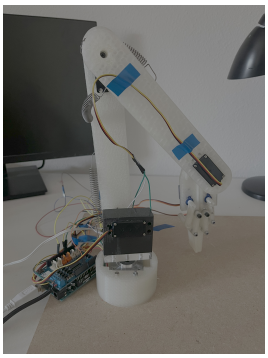

Figure 7: 3D-printed Robot Arm Controlled with Arduino.

A.2 Using Leap Motion for Gesture Recognition

Initially, we planned to use the Leap Motion device for tracking gestures, but this approach was abandoned for several reasons. The additional device complicated the system setup, and in practice it did not provide any clear advantage over a standard camera combined with OpenCV and MediaPipe for gesture tracking. The system proved to be more robust when using MediaPipe, with more reliable gesture recognition overall. When using the Leap Motion device, hands were often not recognized correctly, as users had to keep them in very specific positions for detection. Maintaining these positions for the duration of the experiment was physically challenging. This issue was less evident with the camera positioned in front of the user, which allowed participants a greater range of motion and more natural hand positions. For these reasons, the camera-based approach was ultimately selected.