Embedded vs. Situated: An Evaluation of AR Facial Training Feedback

Avinash Ajit Nargund
Electrical and Computer Engineering
University of California, Santa Barbara
Santa Barbara, California, USA
anargund@ucsb.edu

Andrea M. Park
University of California, San Francisco
San Francisco, California, USA
andrea.park@ucsf.edu

Tobias Höllerer
Computer Science
University of California, Santa Barbara
Santa Barbara, California, USA
holl@cs.ucsb.edu

Misha Sra
Computer Science
University of California, Santa Barbara
Santa Barbara, California, USA
sra@ucsb.edu

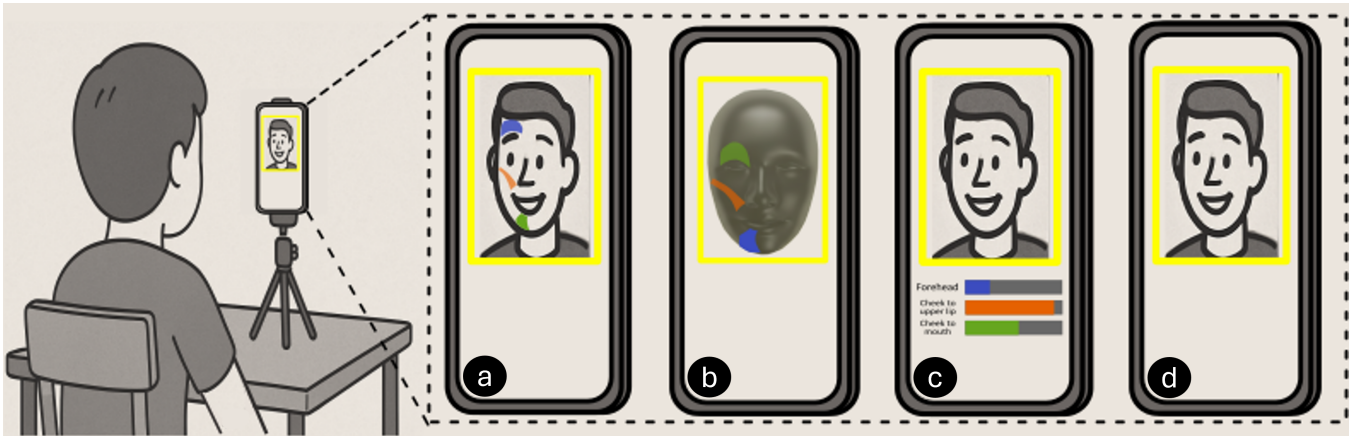


Figure 1: A participant engages in AR-guided facial exercises, receiving real-time feedback designed to improve muscle movement patterns. We evaluate three AR-based feedback conditions, (a) embedded *ARSelfie* with feedback overlaid on the user's own face, (b) proxy-embedded *Mannequin* with feedback on a 3D avatar, (c) situated *BarChart* with feedback conveyed through bars located below the user's face and (d) a no-feedback *Baseline* condition where the user can see themselves. We assessed how the spatial placement of feedback affects user performance, experience and cognitive load. [Portions of this image were generated using AI.]

Abstract

While augmented reality (AR) research demonstrates benefits of embedded visualizations for gross motor training, its applicability to facial exercises remains under-explored. Providing effective real-time feedback for facial muscle training presents unique design challenges, given the complexity of facial musculature. We developed three AR feedback approaches varying in spatial relationship to the user: situated (screen-fixed), proxy-embedded (on a mannequin), and fully embedded (overlaid on the user's face). In a within-subjects study (N=24), we measured exercise accuracy, cognitive load, and user preference during facial training tasks. The embedded feedback reduced cognitive load and received higher preference ratings, while the situated feedback enabled more precise corrections and higher accuracy. Qualitative analysis revealed a key design tension: embedded feedback improved experience but created self-consciousness and interpretive difficulty. We distill these insights into design considerations addressing the trade-offs for facial training systems, with implications for rehabilitation, performance training, and motor skill acquisition.

CCS Concepts

• **Human-centered computing** → **Empirical studies in visualization**.

Keywords

augmented reality, facial muscle visualization, training

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791941

ACM Reference Format:
Avinash Ajit Nargund, Andrea M. Park, Tobias Höllerer, and Misha Sra. 2026. Embedded vs. Situated: An Evaluation of AR Facial Training Feedback. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3772318.3791941

1 Introduction

Facial muscles play a crucial role in human communication, emotional expression, and various physiological functions [45]. Monitoring and training of facial muscle movements can offer benefits in specialized applications. These include actors and performers seeking to refine emotional expressivity [5, 6, 18], public speakers aiming to enhance vocal delivery [15, 21], and patients undergoing rehabilitation for facial paralysis [4] or hypomimia [9]. In these settings, fine-grained guidance is essential to help users isolate and activate specific facial muscles. This work evaluates visualization efficacy independent of specific application. We believe that it is important across application domains to understand visualization impact on 1) exercise accuracy, 2) cognitive and task load, and 3) user preference.

Existing systems for facial muscle training provide limited forms of visual feedback, typically displayed adjacent to the user's face. For example, Farapy [4] overlays bar charts below the user's mirrored face to indicate muscle activation, while GY MEDIC [22] and related systems place raw symmetry scores or text labels beside the face [9]. These screen-adjacent visualizations are spatially disconnected from the region of movement, which may hinder a user's ability to interpret and act on the feedback [4]. While these tools support basic self-monitoring, they fall short of offering spatially integrated guidance that aligns with the movement itself.

In contrast, interactive systems for full-body and gross motor training have extensively leveraged spatially embedded feedback using augmented reality (AR) [2, 49, 52, 61]. These systems commonly align visual cues with the user's body or movement path, using AR mirrors, wearable displays, or head-mounted overlays [48, 52, 60]. Such embedded visualizations [55] enhance user comprehension by co-locating feedback with the physical action. Though well-studied for limb and posture training [47, 55], these techniques remain underexplored for fine-grained and subtle movements of the face.

To address this gap, we investigate how the spatial placement of visual feedback affects user experience and performance during facial muscle training using mobile AR. We evaluate three visualization strategies to convey facial muscle activation that differ in their spatial relationship to the user's face (see Figure 1): (a) screen-anchored *BarChart*, (b) proxy-embedded *Mannequin*, and (c) embedded *ARSelfie* view. Inspired by the visualization used in Farapy [4], *BarChart* conveys muscle activations through proportional bar-fills placed at the bottom of the screen. In the *Mannequin* paradigm, feedback cues are projected on a plain textured 3D model of the user's face while in the *ARSelfie* view the cues are overlaid directly on the user's face. These techniques represent distinct approaches along the "WHERE"-axis of the design space of visual corrective feedback for XR motion guidance systems [59].

We conducted a within-subjects study ($N = 24$) in which participants performed three standardized facial exercises using the three visualization conditions and a no-feedback *Baseline* condition where participants see their face in the camera without any feedback visualization. Our study was guided by the following research questions:

(1) **RQ1:** Does embedded feedback (*ARSelfie* and *Mannequin* conditions) improve (a) exercise accuracy defined as the proportion of muscles activated correctly per repetition, and (b) rate of correctly executed exercise repetitions, when compared with the situated *BarChart* condition or the *Baseline* condition?
(2) **RQ2:** What are the cognitive and task loads associated with each feedback modality when compared to the *Baseline* condition?
(3) **RQ3:** Which condition do users prefer?

Our work makes the following contributions:

- We present three visualization approaches: situated, proxy-embedded, and fully embedded, that extend prior AR feedback work from gross motor to fine-grained facial muscle exercises.
- From a within-subjects study ($N = 24$), we provide empirical evidence that embedded feedback which has been shown to be effective for full-body and gross motor training is also applicable for facial muscle training and results in better user experience compared to situated feedback which promotes precise activation of the facial muscles.
- We synthesize the quantitative and qualitative results to propose design guidelines that inform how future systems can effectively balance exercise performance accuracy, feedback interpretability, and user comfort.

2 Related Work

Our work is related to the use of AR, specifically situated and embedded visualizations in the context of instruction and feedback for facial motor learning. While several works have investigated the AR-based feedback for applications such as gait rehabilitation [44], upper limb training [49] or general physical therapy [17], their effectiveness and suitability for facial exercises remains largely unexplored in prior work. Only a few studies such as Farapy [4], GY MEDIC [22], and Cai et al. [9] have explored the use of AR-based feedback for facial exercises. In this section, we provide a brief overview of prior work to contextualize the contributions of our research.

2.1 Instruction and Feedback for Motor Learning

Acquiring a motor skill is a complex process that depends on effective instruction and feedback. Motor learning involves three sequential stages: cognitive, associative and autonomous [19]. Feedback is most critical during the initial cognitive stage, where the learner is forming a mental model of the movement, and in the associative stage, where they are refining their technique. While learners receive intrinsic feedback through their own sensory systems (e.g., proprioception), augmented feedback provided by an external source, such as a human coach or a computer, is often helpful in accelerating learning and overcoming performance plateaus [42].

One of ways to categorize augmented feedback is by distinguishing between *Knowledge of Results (KR)* which informs the learner about the result of their action (e.g., did they hit the target or not) reducing the *gulf of evaluation* [39], and *Knowledge of Performance (KP)*, which provides information about the correctness and characteristics of the movement itself (e.g., hand was not raised fast enough) which helps diminish the *gulf of execution* [39]. AR-based training systems are capable of sensing user movements in real-time and delivering detailed KP that would be challenging for a human tutor to convey accurately [47].

Our work explores the use of embedded AR visualizations for providing precise, real-time KP for facial exercises. We compare this approach with situated AR visualizations to assess their impact on user performance, experience, and preference.

2.2 Feedback Visualization

Situated visualization, introduced by White [54], refers to the spatial placement of data visualizations close to the physical entities or referents it is associated with. While situated visualizations remain adjacent to the referents, embedded visualizations go further by aligning visuals with the physical extents of the referents [55]. Research in feedback systems for motor skill acquisition and physical rehabilitation has explored a variety of visualization paradigms designed to bridge the gap between a user's current and target state. These systems primarily differ in how the feedback is spatially related to the user's body. The spatial relationship strongly affects the interpretability of the guidance and the cognitive load imposed on the users [47, 55].

Within the broad spectrum of visualization paradigms, prior work in AR-based training systems has used three main strategies: situated, proxy-embedded, and embedded.

Situated feedback. Situated visualizations separate the feedback from the body, displaying it in the surrounding environment, such as projecting foot placements on a treadmill [44], displaying life-size visuals of an expert to help users with rock climbing [29] or providing text-based exercise feedback [11, 16].

Proxy-embedded feedback. Proxy-embedded visualizations map feedback onto avatars or skeleton models scaled to the user's size and body dimensions. YouMove [2] visualized movement guidance on a skeleton displayed on a large AR mirror while MuscleRehab [61] provided feedback by displaying muscle activation levels on a virtual 3D model.

Embedded feedback. Embedded feedback minimizes perception-action gaps by aligning cues directly with the body [55]. LightGuide [48] projected 2D and 3D arrows directly onto user's hands to guide mid-air gestures, improving hand movement accuracy by nearly 85% when compared to video-based guidance while SleeveAR [49] projected color-coded feedback onto a custom sleeve to support shoulder rehabilitation. Some systems use virtual mirrors to overlay visuals on the user's reflection. These systems combine the familiarity of mirrors with augmented feedback to either guide users [52] or highlight errors made by the user [53].

While the efficacy of these paradigms has been demonstrated for upper-limb rehab and gait training, their application to facial exercises, which involve fine motor movement, is underexplored.

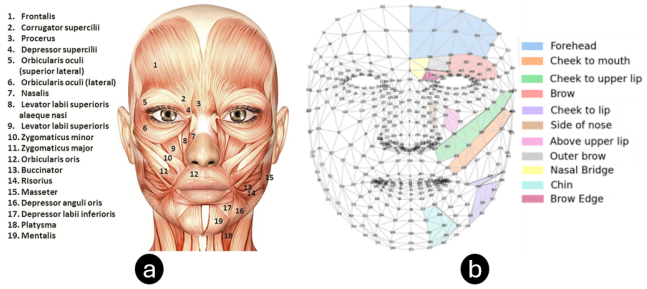


Figure 2: The facial muscle anatomy chart (a) (from [38]) that we used to map subsets of the 468 facial landmarks 3D face mesh (b) provided by Google ARCore *AugmentedFaces* API to the target muscles in our study.

Existing facial exercise systems have predominantly adopted situated visualizations, displaying feedback adjacent to the user's face. For example, Farapy [4] and Cai et al. [9] use AR overlays near the user's selfie view to provide performance feedback. However, the separation of the feedback from the face introduces a disconnect, with Farapy [4] participants reporting that bars at the bottom of the screen made interpreting feedback harder, and suggested that embedded cues, such as color changes directly on the facial muscles, would have been preferable.

In our work, we investigate the effectiveness of three visualization strategies explicitly designed for facial motor training. Using these visualization conditions, we first examine whether embedded visualization, which have demonstrated benefits for limb and whole-body training [59], can be applied to facial exercise training. We then compare how the spatial placement of feedback relative to the face influences accuracy of facial movements, the ease of understanding guidance, and the cognitive load it imposes on the users.

2.3 On-Face Visualization

Prior research has utilized user's reflection to display personal health informatics [1, 13, 25, 40, 50]. Wize Mirror [25] used data from multiple sensors to estimate cardio-metabolic risk factors and display a composite "Wellness Index" next to the user's reflection. Similarly, Rigling et al. [40] explored the use of AR face filters to visualize fitness tracker metrics, such as step count and sleep time.

Building on prior work that visualizes health data on the body or face to reduce the disconnect between data and its context, we extend this approach to facial-muscle training by visualizing muscle activations in real-time on the corresponding facial anatomy to provide users with immediate, contextually grounded feedback.

3 Study Design and Implementation

To investigate our research questions, we developed a mobile AR application[1] using the Unity[2] game engine. The application uses Google ARCore's *AugmentedFaces* Unity API [3] for facial tracking.

[1]Source code available at : https://github.com/HAL-UCSB/embedded_vs_situated
[2]https://unity.com/
[3]https://developers.google.com/ar/develop/unity-arf/augmented-faces/developer-guide

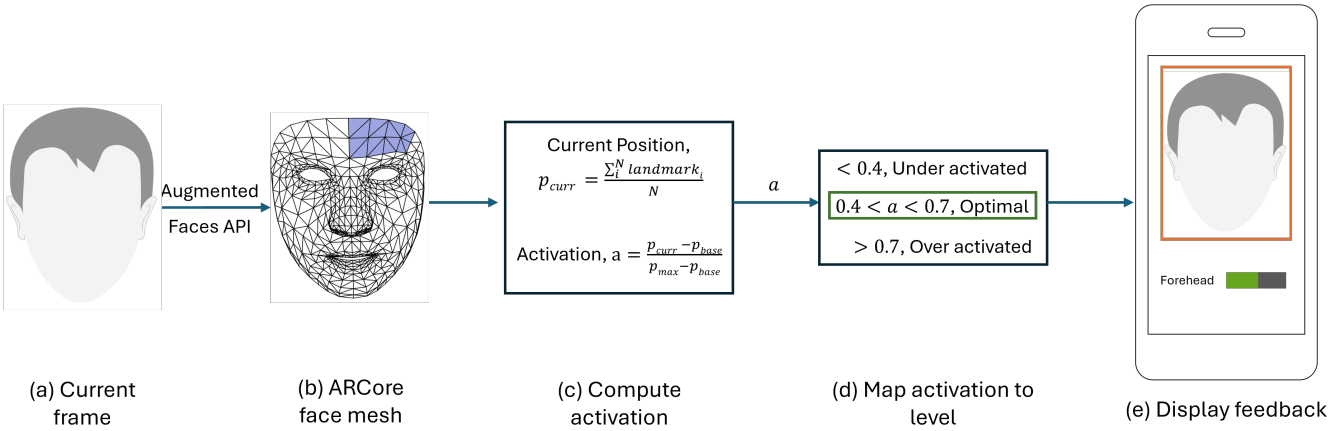


Figure 3: The activation estimation process, demonstrated with the left forehead muscle in the *BarChart* condition. First, the phone's front camera captures the user's face and the *AugmentedFaces* API returns a face mesh. The landmarks corresponding to a particular muscle are used to estimate its current position. This position along with the muscle's baseline and peak positions are used to compute a raw activation score. This score is mapped to an activation label and both are used to provide feedback to the user.

This API provides a 468-point 3D face mesh (Figure 2b) and tracks the face in real-time using the front-facing camera of the phone. We leverage this dense face mesh to calculate muscle activation and to spatially register feedback overlays onto the user's face. The landmark-based approach offers two key advantages over using raw image pixels. First, the pre-trained machine learning (ML) models in ARCore are optimized for mobile inference and trained on large-scale datasets, making them robust to variations in lighting, head pose, and user appearance. Second, the 3D mesh provides a semantically organized and anatomically grounded representation of the face. This directly facilitates derivation of interpretable muscle activation estimates by measuring localized landmark displacements, whereas pixel-based methods often rely on indirect inference and require large, labeled datasets to handle variations in appearance, expression, and lighting.

We design and implement four experimental visualization conditions : *ARSelfie* (embedded), *Mannequin* (proxy-embedded), *BarChart* (situated) and *Baseline*. In this section, we describe our design rationale, system architecture, and features of each visualization condition.

3.1 Mapping Landmarks to Muscles

To provide muscle-specific feedback, we manually map subsets of the 468 ARCore facial landmarks to their corresponding facial muscles. We leverage the biomechanical principle that facial muscles, unlike most other muscles, are attached directly to the skin rather than bone [12]. This direct attachment means that muscle contraction causes localized skin deformation, enabling the use of facial landmark clusters as a proxy to estimate muscle activations [46, 58]. This method is similar to the approach in prior work by Hasebe et al. [24], which establishes that landmark cluster movements correlate with integrated electromyography (iEMG) estimates of muscle activations, particularly for the mouth (significant correlation) and eye muscles. By overlaying the ARCore face mesh onto facial musculature anatomy charts [35], we visually identified clusters of facial landmarks situated over particular muscles, whose displacement indicates activation. The accuracy of the mapping was reviewed and validated by our second author, who is a facial nerve surgeon.

3.2 Participant Calibration

In the calibration phase, the user is first prompted to hold a neutral facial expression, during which the system records the base position of all the facial landmarks. Next, the user performs each target exercise with maximum effort (e.g., smiling as wide as possible) and the system records the 3D landmark positions at peak activation. The maximum movement range of each muscle is then calculated as the 3D Euclidean distance between the centroid of its landmark cluster at peak activation and in the neutral expression. This calibration establishes a personalized reference scale for determining activation levels.

3.3 Muscle Activation

During an exercise, the system estimates muscle activation by tracking the displacement of specific facial landmark clusters (see Figure 3). The centroid of each cluster is used as the position of the corresponding muscle, and the system continuously measures displacement as the Euclidean distance between its current position and its initial position in a neutral expression. The muscle activation is then computed as the ratio of current displacement to maximum movement range established in the calibration step. The normalization accounts for inter-individual variability, as it creates a personalized 0 to 1 scale for all the muscles of each user based on their range of motion. The activation score is then mapped to a three-class label. Scores below 0.4 are classified as under-activation, scores above 0.7 as over-activation and scores in between as optimal. These thresholds were determined through pilot testing and

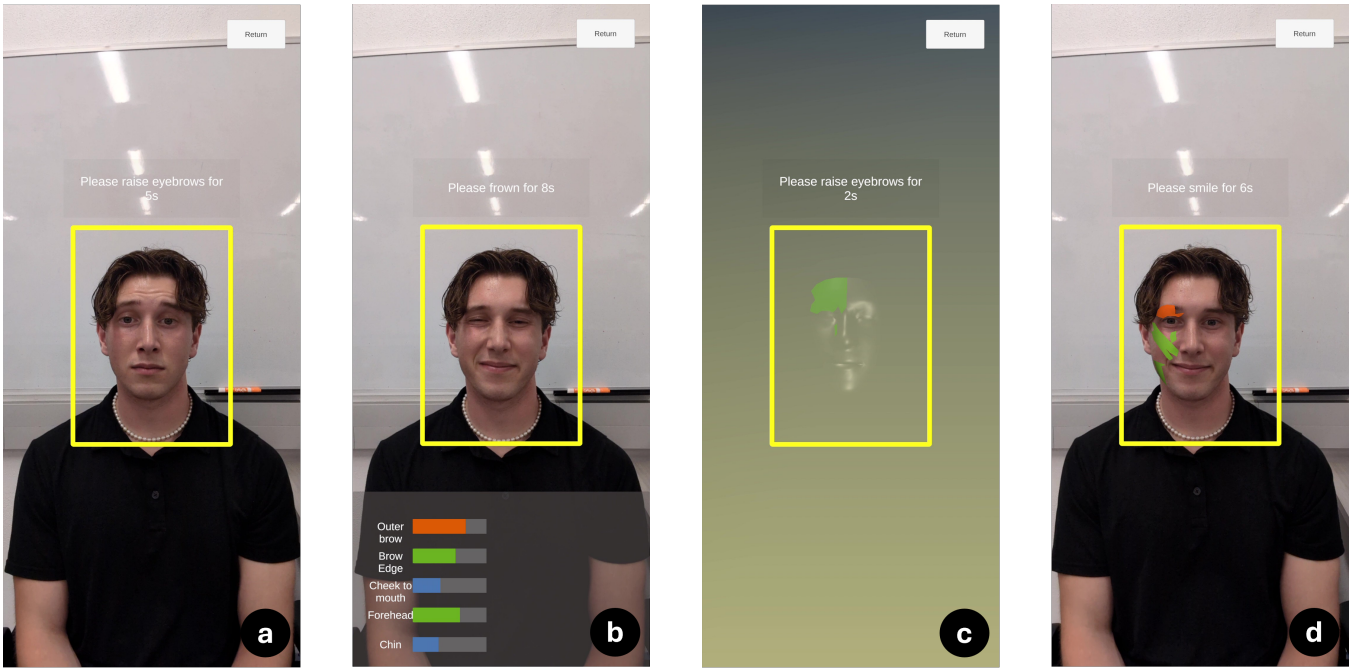


Figure 4: The four experimental conditions used in the study, showing a participant performing the three facial exercises : (a) *Baseline* (*Eyebrow Raise*), (b) *BarChart* (*Reverse Frown*), (c) *Mannequin* (*Eyebrow Raise*) and (d) *ARSelfie* (*Smile*).

consultation with a facial surgeon. The three-way classification is similar to the approach used in prior work [4].

For the study, we use a landmark displacement-based heuristic instead of a machine learning (ML) model to ensure deterministic and stable muscle activation estimates. Controlling the variability at the activation estimation level enables us to isolate the effects of the visualization design and avoid potential confounds from noisy or inaccurate ML predictions. Additionally, this heuristic approach minimizes computation load on the mobile device, ensuring the feedback interface is fast and responsive.

3.4 Visualization Conditions

To explore how the spatial location and level of embodiment of the feedback impacts users, we designed four visualization conditions (see Figure 4).

Baseline. In the *Baseline* condition, the user's live camera feed is displayed with no activation score-based feedback.

Situated Visualization: BarChart. The *BarChart* condition provides muscle activation feedback through horizontal bar charts placed at the bottom of the device screen, detached from the user's live camera feed shown at the top. Each bar corresponds to a target muscle in the exercise and is labeled using the muscle's common name (e.g.,*Frontalis* muscle is labeled "Forehead"). The width of the bar fill is proportional to the muscle's raw activation, which provides continuous feedback on the activation intensity to the user. The bar fills are colored based on the activation level: blue for under-activated, green for optimally activated and orange for over-activated.

Proxy-Embedded Visualization: Mannequin. The *Mannequin* condition provides visual feedback on a dynamic 3D facial avatar created using the ARCore face mesh. The avatar replicates the user's facial movements in real time via mesh deformation but omits photorealistic textures, appearing instead as a neutral proxy face. Feedback is conveyed through colored overlays aligned with specific muscle regions, generated by segmenting the mesh into sub-regions that correspond to the targeted muscles.

This visualization condition uses both color and opacity to communicate feedback. This approach is consistent with the method used by Ikeda et al. [26] and aligns with the "Decal" pattern for situated visualizations described by Lee et al. [33] (see Figure 5). The color of an overlay represents the discrete activation level: blue for under-activated, green for optimal, and orange for over-activated. The opacity of the overlay is indicative of the absolute difference between the current and optimal activation score. In the under-activated range (scores below 0.4), the blue overlay's opacity decreases as the activation approaches the optimal threshold, encouraging the user to activate their muscles further. In the optimal range ($0.4 - 0.7$), the green overlay is most opaque at the center of the range (a score of ~ 0.55) and becomes gradually more transparent toward the lower and upper bounds of optimality. In the over-activated range (scores above 0.7), the orange overlay's opacity increases as the activation score increases, indicating excessive activation.

Embedded Visualization: ARSelfie. The *ARSelfie* condition uses the same design, feedback mechanism and colors as the *Mannequin*. The main difference is that the feedback is overlaid directly onto

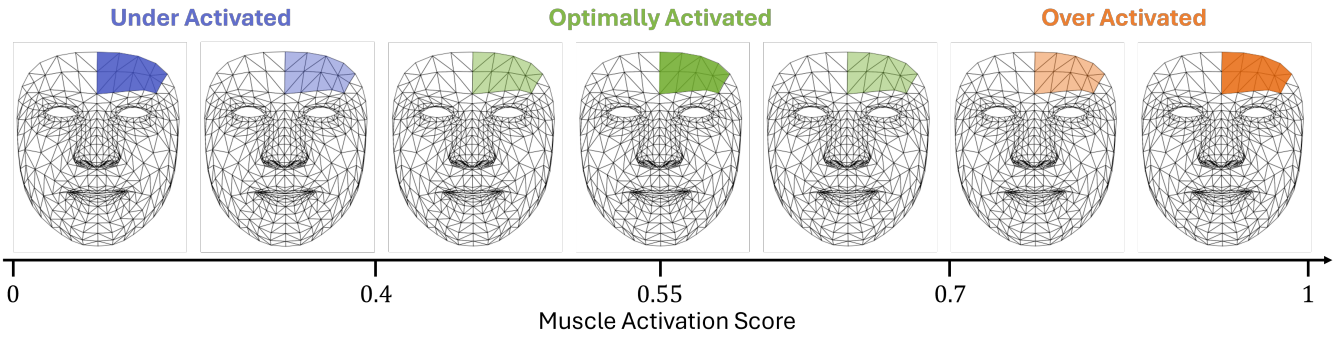


Figure 5: Opacity-modulated feedback in the embedded conditions, illustrated for the forehead muscle on the ARCore face mesh. Color indicates the activation zone: blue for under-activation, green for the optimal range, and orange for over-activation. Opacity encodes distance from the optimal range: the blue overlay fades as activation approaches green, the green overlay is most opaque at the midpoint and more transparent near its lower and upper thresholds, and the orange overlay increases in opacity as activation rises further beyond optimal.

the user's face in the live video feed rather than on a proxy avatar face.

3.5 Data Logging

The mobile application logs the raw activation scores for all target muscles and identifiers for the current exercise into separate files for each condition. This data is captured at the application's frame rate (synchronized to Unity's `Update()` cycle) and is used to compute performance metrics including the proportion of exercise time during which muscles were optimally activated, the time taken to achieve the first optimal repetition, and the total number of optimal repetitions within the exercise window.

4 User Evaluation

We conducted a user study to examine how different spatial placements of visual feedback for facial motor training influence user performance, perceived task load and experience as they perform facial exercises. To run the study, we developed a custom application that provided three types of AR-based feedback, *ARSelfie*, *Mannequin* and *BarChart*, as well as a no-feedback *Baseline* (described in Sec. 3). The study was conducted in a quiet, indoor room with consistent lighting. The application ran on a Google Pixel 8 Pro smartphone, placed 0.5 meters in front of the participant. A chin rest was used to stabilize the participant's head and minimize face tracking inconsistencies from unintended head movements. The facial exercises used in the study, study procedure (outlined in Figure 6) and the task are described below.

4.1 Facial Exercises

Participants performed three facial exercises, smile, eyebrow raise, and reverse frown, each requiring the activation of muscles of different sizes from distinct facial regions (See Figure 4). The smile relied on the engagement of muscles around the mouth. The eyebrow raise involved raising the eyebrows toward the hairline which is mainly controlled by the forehead muscle. The reverse frown, commonly known as "face scrunch," required participants to draw their eyebrows down and together while moving the muscles in their cheeks and chin upwards. This compound movement requires coordinating muscle activation across the upper and lower face.

4.2 Study Procedure

Participants began the study by providing informed consent (study approved by our local IRB under protocol #23-25-0400), demographic information and completing the Affinity for Technology Interaction (ATI) questionnaire [20], which assesses individual differences in interaction and engagement with technology. Participants were briefed on the purpose of the study and introduced to the three facial exercises through a tutorial. The tutorial also explained the visualization encodings including the color scheme and opacity modulation used across conditions. Participants also learned that optimal repetitions are indicated by muscle-feedback indicators turning green for all muscles for that exercise.

Following the tutorial, participants completed a calibration phase. In this phase, they were first instructed to maintain a neutral expression for 10 seconds to record their baseline facial landmark positions. Next, they performed the three exercises with maximal effort with a 5 second break between exercises. This calibration also helped familiarize participants with the exercises they would perform in the exercise phase (detailed in Section 4.3).

During the exercise phase, after each visualization condition, participants completed the NASA-TLX questionnaire [23], the User Experience Questionnaire (UEQ) [32] and a custom questionnaire adapted from [28] to assess extraneous cognitive load (ECL) (Table 1). Measuring the extraneous cognitive load allows us to estimate how demanding each visualization condition made it for the participants to extract task-relevant information and perform the exercises effectively. At the end of the study, participants completed a post-study questionnaire where they ranked the conditions in order of preference and provided rationale for their preferences in a semi-structured interview with the researcher.

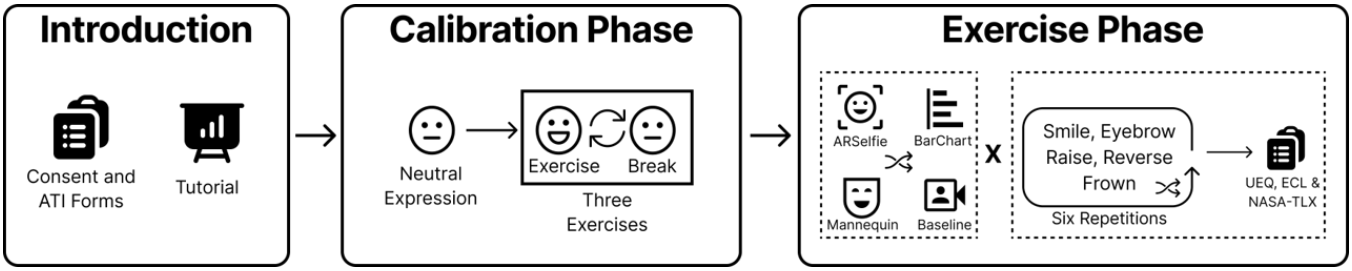


Figure 6: Study procedure. After providing informed consent and completing the ATI questionnaire [20], participants received a tutorial on exercises, study task and feedback conditions through a presentation, after which they switched to the system for the calibration and exercise phases. In the calibration phase participants first maintained a neutral expression (10 seconds) and then performed the three exercises at maximum effort (10 seconds each with 5-second breaks). In the exercise phase, participants completed four counter-balanced trials (Latin-square design), one per condition. Each trial consisted of six repetitions of the exercise cycle (three exercises performed for 10 seconds in randomized order with 5-second breaks). After each trial, participants completed the NASA-TLX [23], UEQ [32], and a custom extraneous cognitive load questionnaire.

4.3 Study Task

During the exercise phase, participants completed four trials, one per visualization condition. Trial order was counterbalanced using a Balanced Latin Square to minimize learning effects and reduce potential carryover between conditions. Each trial consisted of six sets, and each set contained all three exercises with their order permuted to mitigate sequence effects. For every exercise, participants worked for 10 seconds followed by a 5-second rest. They were instructed to perform as many optimal repetitions as possible within each 10-second window, where a repetition was counted as optimal only when all muscle-feedback indicators appeared green. The exercise and break durations were determined through pilot testing to provide sufficient time for participants to perform multiple repetitions per exercise window while minimizing fatigue over the full study session.

4.4 Participants

Following initial pilot tests, a power analysis was conducted to determine the required sample size. Assuming small-to-medium effect size ($f = 0.2$), significance level of $\alpha = 0.05$, a statistical power of $1 - \beta = 0.8$ and correlation of 0.7 among repeated measures, the analysis indicated a target sample of 22 participants. We subsequently recruited 24 participants (10 female, 14 male) aged between 18 to 30 (median = 20) years using internal mailing lists. All participants had normal or corrected-to-normal vision. Seventeen participants reported no color vision deficiencies. Two participants self-reported minor color vision deficiencies but confirmed that they could distinguish between colors used in the study. The remaining five participants did not disclose their color vision status but reported no difficulty distinguishing colors before or during the study. All participants were compensated $15 for their time.

5 Quantitative Results

We evaluated the following dependent variables to assess the effect of feedback conditions: (a) performance metrics including exercise accuracy, time to first optimal repetition and number of repetitions, (b) user experience ratings measured through the User Experience Questionnaire (UEQ) [32], (c) task load measured via the NASA-TLX [23], and (d) extraneous cognitive load (ECL) scores.

A Shapiro-Wilk test was conducted on all dependent variables to assess normality. We used a repeated measures ANOVA when the assumption of normality was met and a Friedman test otherwise. All post-hoc pairwise comparisons were performed using Bonferroni-Holm correction.

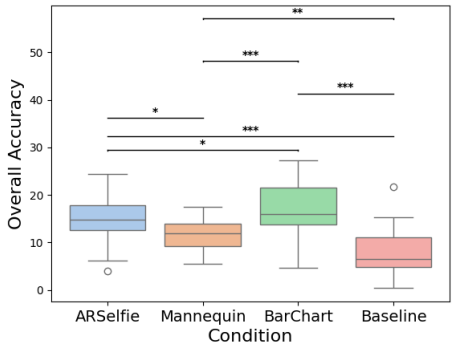
5.1 Exercise Performance

We measured the exercise performance using three objective metrics :

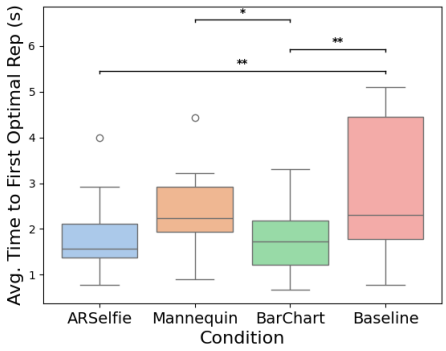
(1) **Accuracy**, the proportion of total exercise time the muscles were optimally activated, averaged across all exercise windows.
(2) **Number of optimal repetitions**, the average number of optimal repetitions completed within each 10-second exercise window.
(3) **Time to first optimal repetition**, the time in seconds from the start of the exercise until the first optimal repetition was completed.

The average performance metrics for each visualization condition are presented in Figure 7.

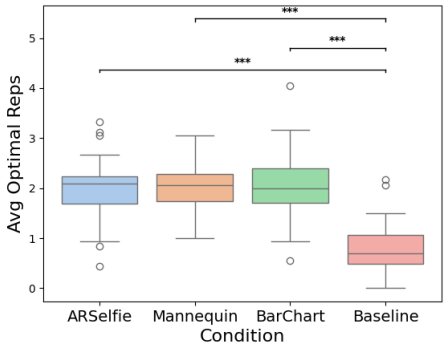
Accuracy. We found a significant main effect of the visualization condition on the average accuracy, $F_{3,69} = 26.16, p < 0.001$. Post-hoc comparisons revealed that the participants performed significantly better in the *BarChart* ($\mu = 16.88, \sigma = 5.7$) condition compared to *ARSelfie* ($\mu = 14.59, \sigma = 5.3, p = 0.046$), *Mannequin* ($\mu = 11.49, \sigma = 3.18, p < 0.001$) and *Baseline* ($\mu = 7.97, \sigma = 5.1, p < 0.001$) conditions. Interestingly, although the *Mannequin* condition is conceptually similar to the *ARSelfie*, accuracy was significantly lower in the *Mannequin* condition ($\mu = 11.49$) compared to *ARSelfie* ($\mu = 14.59, p = 0.019$).



(a) Activation accuracy

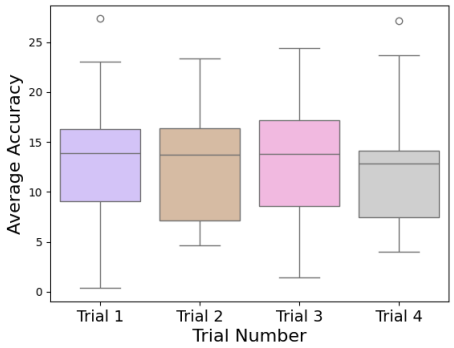


(b) Time to first optimal repetition

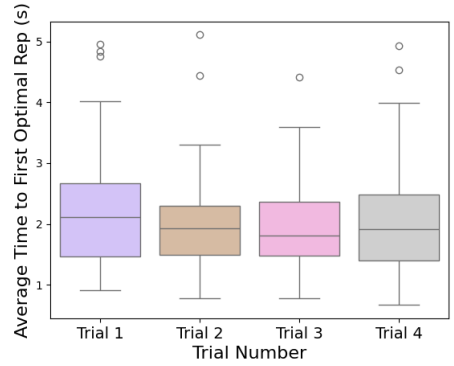


(c) Number of repetitions

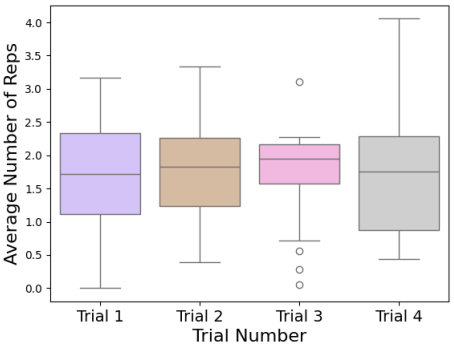
Figure 7: The performance of the participants is measured using (a) average accuracy, (b) time to first optimal repetition and (c) number of optimal repetitions. Participants were significantly more accurate with the *BarChart* visualization. The time to first optimal repetition in the *ARSelfie* and *BarChart* conditions was significantly shorter compared to the *Baseline*. Participants performed significantly fewer repetitions in the *Baseline* condition compared to the visualization conditions. The whiskers indicate significant post-hoc comparisons (* $p < 0.05$; ** $p < 0.01$, *** $p < 0.001$). All subsequent plots with statistical test results follow the same notation.



(a) Activation accuracy (trials 1-4)

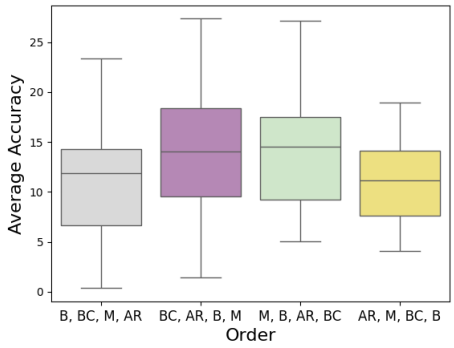


(b) Time to first optimal repetition (trials 1-4)

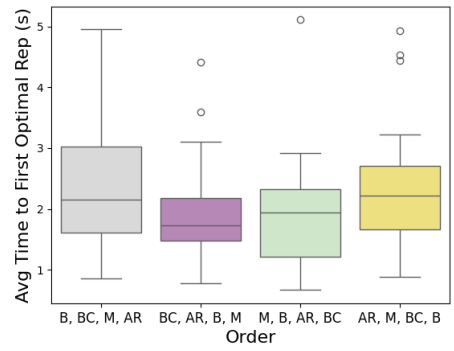


(c) Number of repetitions (trials 1-4)

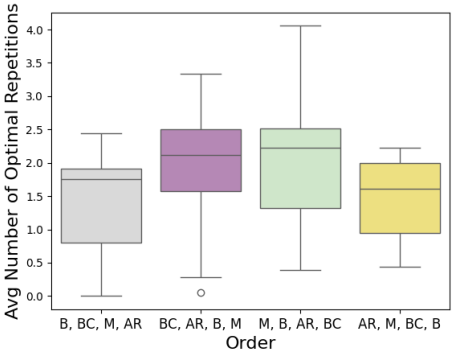
Figure 8: The average accuracy, time to first optimal repetition and number of optimal repetitions by trial number (1-4) to assess for learning effects. The plots show that participant performance on all metrics remained stable from the first trial to the last, indicating that no significant learning effects occurred during the experiment.



(a) Activation accuracy across the four condition orders



(b) Time to first optimal repetition across the four condition orders



(c) Number of repetitions across the four condition orders

Figure 9: Performance metrics for each of the four presentation order groups to assess for carryover effects. There were no significant differences in performance between the groups for any of the three metrics. This suggests that the order in which participants viewed the visualizations did not significantly influence their performance.

Number of optimal repetitions. A repeated-measures ANOVA showed significant main effect of the visualization on the average number of optimal repetitions, $F_{3,69} = 35.9, p < 0.001$. Post-hoc analysis revealed that the participants performed significantly more repetitions in the *ARSelfie* ($\mu = 1.97$), *Mannequin* ($\mu = 2.01$), and *BarChart* ($\mu = 2.08$) conditions than in the *Baseline* condition ($\mu = 0.81$; all $p < 0.001$).

Time to first optimal repetition. A Friedman test showed a significant main effect of visualization on the average time to first optimal repetition, $Q = 13.1, p = 0.004, W = 0.18$. Post-hoc pairwise comparisons indicated that participants took significantly more time to make their first optimal repetition in the *Baseline* condition ($\mu = 2.85s, \sigma = 1.46s$) when compared to the *ARSelfie* ($\mu = 1.76s, \sigma = 0.7s, p = 0.003$) and *BarChart* ($\mu = 1.7s, \sigma = 0.65, p = 0.005$) conditions. Notably, users were also significantly slower in the *Mannequin* condition ($\mu = 2.32s, \sigma = 0.78s$) when compared to the *BarChart* ($p = 0.01$).

5.1.1 Learning and Carryover Effects. We analyzed the performance data to determine if participant performance improved as they completed more trials or if the (counterbalanced) order of conditions influenced the results.

To test for learning effects, we assessed performance across the four trials (shown in Figure 8). A repeated-measures ANOVA test show no significant effect of the trial number on either the average accuracy, $F_{3,69} = 0.19, p = 0.9$ or average number of optimal repetitions, $F_{3,69} = 0.04, p = 0.98$. Similarly, a Friedman test revealed no significant effect of the trial number on the average time to first optimal repetition, $Q = 2.2, p = 0.53, W = 0.03$.

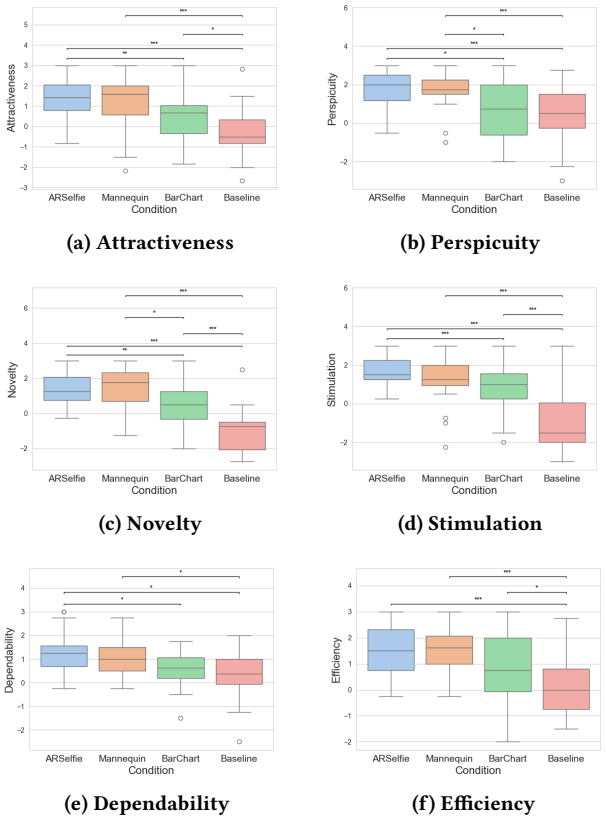
To test for carryover effects, we used a mixed-ANOVA test with condition as within-subject factor and the visualization presentation order as the between subject factor (shown in Figure 9). The analysis showed no significant effect on accuracy ($F = 1.55, p = 0.23, \eta_p^2 = 0.19$) or number of optimal repetitions ($F = 2.59, p = 0.08, \eta_p^2 = 0.28$). For non-normal distributions we used an Aligned Rank Transform ANOVA [56], which also revealed no significant effect of the order, $F = 2.22, p = 0.09, \eta_p^2 = 0.07$. This indicates carryover effects being negligible.

5.1.2 Effect of ATI and Familiarity with AR. We conducted a correlation analysis to determine if the individual differences in ATI [20] scores or familiarity with AR influenced the performance. The analysis revealed no significant correlations between either the ATI scores or AR familiarity and any of the three performance metrics.

5.2 User Experience

The user ratings across the six scales of the User Experience Questionnaire (UEQ) [32] for the four conditions is shown in Figure 10. Overall, the *ARSelfie* and *Mannequin* conditions consistently scored highest across both the pragmatic and hedonic aspects of user experience. The *BarChart* condition was generally rated higher than *Baseline* but lower than the two embedded conditions.

Pragmatic quality refers to the task-related aspects of the user experience. It is composed of the *Perspicuity*, *Efficiency*, and *Dependability* sub-scales. For *Perspicuity*, a Friedman test showed a significant main effect of the visualization condition ($W = 0.35, p <$



(a) Attractiveness (b) Perspicuity

(c) Novelty (d) Stimulation

(e) Dependability (f) Efficiency

Figure 10: The six sub-scales of the User Experience Questionnaire [32] for each of the experimental conditions. The *ARSelfie* and *Mannequin* conditions were consistently rated higher than *BarChart* and *Baseline* on both pragmatic and hedonic dimensions. Higher is better with scores > 0.8 considered positive and < -0.8 considered negative [43].

0.001). Post-hoc comparisons indicated that the *ARSelfie* and *Mannequin* conditions were perceived as significantly clearer than *BarChart* ($p = 0.019$ and $p = 0.016$, respectively) and *Baseline* ($p < 0.001$ for both).

For *Efficiency*, a Friedman test revealed a significant main effect ($W = 0.5, p < 0.001$). Post-hoc comparisons showed that the participants perceived the *Baseline* condition as significantly less efficient in guiding them than *ARSeflie* ($p < 0.001$), *Mannequin* ($p < 0.001$), and *BarChart* ($p = 0.026$). No significant differences were found among the other three conditions.

For *Dependability*, a repeated measures ANOVA found a significant main effect ($F_{3,69} = 8.41, p < 0.001, \eta_g^2 = 0.16$), with a Greenhouse-Geisser correction applied as the sphericity assumption was violated. Post-hoc tests showed that the participants found the *ARSelfie* condition more dependable than both *BarChart* ($p = 0.023$) and *Baseline* ($p = 0.012$). The *Mannequin* visualization was also rated as more dependable than the *Baseline* condition ($p = 0.011$).

Hedonic quality measures the affective aspects of the interaction. It consists of the *Stimulation* and *Novelty* sub-scales.

For *Stimulation*, a Friedman test indicated a significant main effect ($W = 0.65, p < 0.001$). The *Baseline* condition was rated as significantly less stimulating than the *ARSelfie* ($p < 0.001$), *Mannequin* ($p < 0.001$) and *BarChart* ($p < 0.001$). Moreover, the *ARSelfie* was rated more stimulating than *BarChart* ($p < 0.001$).

For *Novelty*, a Friedman test revealed a significant main effect ($W = 0.62, p < 0.001$). Post-hoc comparisons indicate a definite ordering, with *ARSelfie* and *Mannequin* rated as more novel, followed by *BarChart* with *Baseline* being rated as least novel. All the pairwise comparisons along this ordering except the one between *ARSelfie* and *Mannequin* were significant ($p < 0.05$).

The *Attractiveness* scale provides a global measure of each condition's overall appeal. A repeated measures ANOVA found a significant main effect ($F_{3,69} = 15.63, p < 0.001, \eta_g^2 = 0.24$). The *Baseline* condition was rated as significantly less attractive than *ARSelfie* ($p < 0.001$), *Mannequin* ($p < 0.001$) and *BarChart* ($p = 0.014$). Additionally, *ARSelfie* was rated as more attractive than *BarChart* ($p = 0.008$).

5.3 Task Load

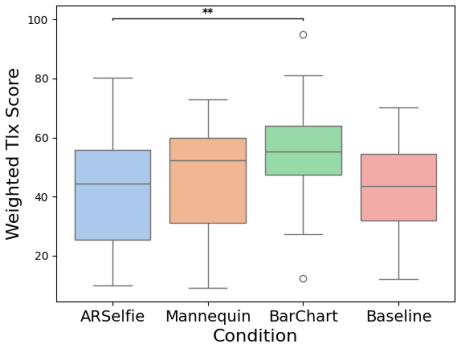
A repeated-measures ANOVA was conducted to assess the impact of the visualization condition on the weighted overall NASA-TLX workload score. As the data violated the assumption of sphericity a Greenhouse-Geisser correction was applied. The analysis revealed a significant main effect, $F_{3,69} = 4.6, p = 0.011, \eta_g^2 = 0.083$. Post-hoc pairwise comparisons using Bonferroni-Holm corrections showed that the workload for the *ARSelfie* condition ($\mu = 42.8, \sigma = 18.6$) was significantly lower than the *BarChart* condition ($\mu = 55.7, \sigma = 17.6, p = 0.009$). The overall workload in the *Baseline* condition ($\mu = 43.19, \sigma = 16.1$) was lower than in the *BarChart* ($\mu = 55.7, \sigma = 17.6$) and *Mannequin* ($\mu = 46.5, \sigma = 17.9$) conditions. However, the pairwise differences were not significant. The results are shown in Figure 11a.

To further investigate the sources of task load, we analyzed the six individual TLX sub-scales. The results are shown in Figure 12.

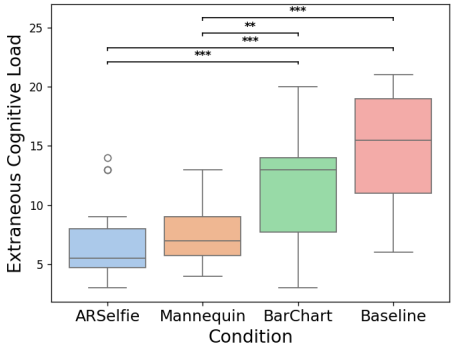
5.4 Extraneous Cognitive Load

Extraneous cognitive load (ECL) is caused by the design of the learning material or application [28, 51]. Users often have to invest cognitive resources in extraneous processes such as searching for information, ignoring irrelevant information while learning or performing a task which might impair their efficiency [51]. We assessed ECL using a modified version of the questionnaire used in Klepsch et al. [28]. It consists of three 7-point Likert-scale questions (see Table 1) with Cronbach's $\alpha = 0.82$. The overall ECL is computed as the sum of the three item scores. The results are shown in Figure 11b.

A Friedman test revealed a significant main effect of visualization condition on the ECL, ($Q = 21.78, W = 0.49, p < 0.001$). Post-hoc comparisons showed that participants reported significantly lower ECL in the *ARSelfie* ($\mu = 6.50, \sigma = 3.16$) and *Mannequin* ($\mu = 7.71, \sigma = 2.53$) conditions compared to both the *BarChart* ($\mu = 11.42, \sigma = 4.77$; $p = 0.001$ and $p = 0.008$, respectively) and *Baseline* ($\mu = 14.88, \sigma = 4.79$; $p < 0.001$ for both). Additionally,



(a) Weighted NASA-TLX Score.



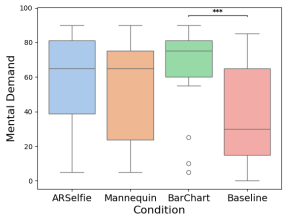
(b) Extraneous Cognitive Load

Figure 11: (a) The weighted NASA-TLX scores for each of the conditions (lower is better). Post-hoc comparisons showed a significantly higher task load in the *BarChart* condition when compared to *ARSelfie*. (b) The ECL scores for the four visualization conditions (lower is better). The ECL associated with embedded conditions were significantly lower than the *BarChart* and *Baseline* conditions.

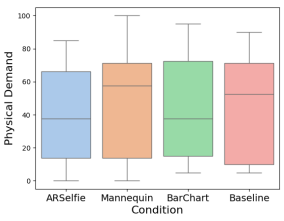
Table 1: The questionnaire used to assess extraneous cognitive load. Responses were captured on a 7-point scale (1=*Strongly Disagree*, 5=*Strongly Agree*). The items were adapted from the [28] to assess how hard it was for participants to obtain the information needed from the visualization to activate their muscles optimally.

1.	It was exhausting to find the muscle activation information needed from this visualization
2.	This visualization is inconvenient for verifying if the muscles are activated optimally
3.	This visualization made it difficult to recognize information and link it to the task

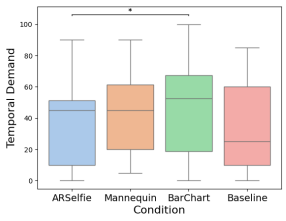
Mannequin was rated as imposing significantly less ECL than *BarChart* ($p = 0.008$). No significant difference was observed between the embedded conditions.



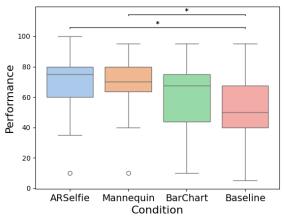
(a) Mental Demand



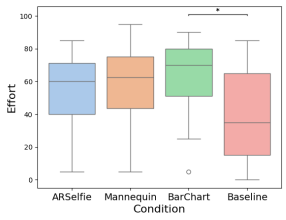
(b) Physical Demand



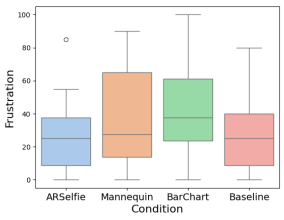
(c) Temporal Demand



(d) Performance



(e) Effort



(f) Frustration

Figure 12: The NASA-TLX sub-scale scores for each of the visualization conditions. Lower scores are better for all scales except *Performance*. The perceived mental demand, temporal demand and effort for *BarChart* condition was higher than all other conditions.

6 Qualitative Results

6.1 Open-ended questions

After each visualization condition, participants provided written responses to the question, *"What did you like about this visualization."* After the whole study they provided written responses to the question, *"Which visualization did you prefer overall? Why?."* Finally, they provided open-ended feedback about the different visualization conditions in a semi-structured interview with the researcher. Participants were specifically asked to elaborate on the aspects of the visualizations that they liked and why they preferred one condition over the other. Their responses to these open-ended questions were coded and organized into themes using inductive thematic analysis [7].

Theme 1: Cognitive and Task Load. Participants frequently noted that both the location of the feedback and its visual encoding affected the amount of effort required to incorporate it.

Visualization Location. The *BarChart* condition required frequent gaze shifts between the bars and the face which many found effortful, with P6 mentioning, "looking down and up made it harder

to assess the state" and P23 was "annoyed a little bit about having to look up and down". Others found these gaze shifts distracting, such as P10 who "could not concentrate on the face and kept looking at the feedback" and P11 who "liked it (*BarChart* condition) the least as I had to look in two different places". On the other hand, the co-located feedback in the embedded conditions reduced gaze shifts. P24 in the *ARSelfie* condition mentioned that they, "didn't have to switch [their] attention to different parts of the screen" and P7 in the *Mannequin* condition highlighted they only needed to "pay attention to one region."

Visual Encoding. Participants found the proportional bar fills with color changes in the *BarChart* condition provided clear, quantifiable feedback. Three out of 24 participants (P14, P22, P23) noted that the *BarChart* showed them "how much [they] had to move". The opacity-based encodings in the embedded conditions were considered less-intuitive. While P5 found them useful for tracking activation five out of 24 participants (P8, P14, P18, P22, P26) described them as more difficult to interpret, requiring more concentration. P18 explained that it was "more effortful to concentrate on the opacity [and] easier to focus on the left/right of the bar chart" while P14 summarized that "transparency was a lot harder to interpret than the bar fill."

Theme 2: Perceived Performance and Learning. Participants pointed out different strengths of each visualization.

Learning. Seven out of 24 participants (P7, P14, P17, P18, P19, P24, P26) perceived the embedded conditions as better for learning how to control the facial muscles. P7 found that *ARSelfie* condition made "it a little easier to learn how much force I need to apply over time". Both P19 and P24 mentioned that seeing their own facial form in the *ARSelfie* condition made it easier to remember correct movements. Even though P18 preferred another visualization overall (*BarChart*), they stated that the *Mannequin* condition "made it very easy to learn which muscle needed to move in what way" while P26 thought that the *Mannequin* condition had the "clearest approach to showing muscle movement that I could learn and adjust from."

In contrast, participants reported limited learning in the *BarChart* condition, because either they often ignored the muscle labels or could not map them to regions of their face. P17 admitted that they "didn't even read the labels and just looked at the colors" and they "[were] not learning at all". Similarly, P24 mentioned that they "didn't read the labels" so they "could not figure out the correspondence [between the feedback and facial muscles]". Some participants despite reading the muscle labels struggled with the mapping. P3 mentioned being "not able to figure [out] what particular part had to be changed" and P26 similarly stated that "it was hard to correlate the muscle names to specific regions of the face". Even when the mapping was understood, the additional effort required for the attention shifts reduced the utility of the feedback and hindered learning. P6 recalled that "mapping feedback to movement took time and brain power," which may have led to some participants (e.g., P21) to stop observing their face and focus only on the feedback.

Performance. Many felt that the *BarChart* condition was effective for immediate performance. P18 noted that the proportional bar fills eliminated "ambiguity from the color/opacity system" and enabled

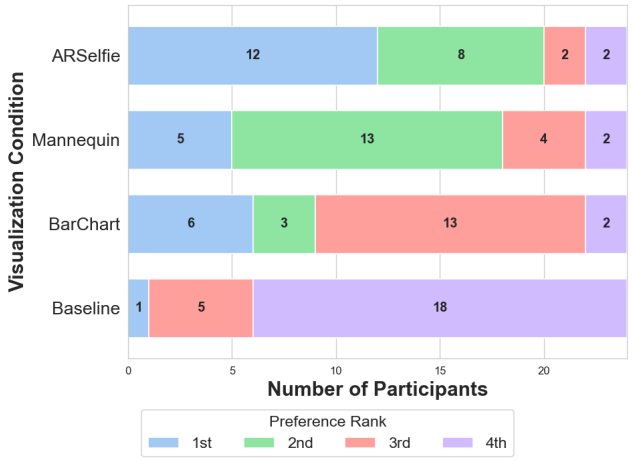


Figure 13: User preference rankings for the four visualization conditions. *ARSelfie* condition received the highest number of first-place rankings, while the *Mannequin* was most frequently ranked second. *BarChart* was most often ranked third, and *Baseline* was the least preferred condition, receiving the highest number of fourth-place rankings.

"to tweak things much more easily". P14 stated they "liked being able to see how close to optimal activation" they were and observing "how much [their] muscle movements changed the score". On the other hand, the feedback overlays in the embedded conditions apparently made adjustment harder for some participants. P21 mentioned that the "filters made it slightly harder to know" if their muscle movements were correct and P22 voiced that they "found it hard to correlate what my face was like because of the overlay"

Theme 3 : User Experience and Preference. Many participants preferred the *ARSelfie* condition as they found it easier to relate their movements with their facial form (P3, P5, P19, P20) or simply because they preferred looking at their own face (P23, P24) instead of the avatar. But several (P4, P6, P21, P26) participants reported discomfort, either from self-consciousness or distractions. P21 mentioned they "didn't like looking at [their] face and ... ended up focusing on the features" and P4 stated that it was "distracting to look at the face and did not want to look at the face for long".

For some of the participants the *Mannequin* condition served as a distraction-free alternative to the *ARSelfie*. P4 mentioned that "it eliminated the background and made it easier to just focus on the facial expressions". P6, P9, P21 all expressed that it had "no distractions" and made it easier to focus on the exercises. However, a few participants found the *Mannequin* condition "unnatural" (P11, P19), "less interactive" (P23) and even "creepy" (P20).

6.2 Preferences

The aggregate rankings (shown in Figure 13) from most preferred to least preferred were: *ARSelfie*, *Mannequin*, *BarChart* and *Baseline*. The *ARSelfie* condition was the most preferred, with 20 out of 24 participants ranking it in their top two, followed by the *Mannequin* condition (18 of 24). In contrast, the *BarChart* (9 of 24) and

Baseline (6 of 24) conditions were ranked in the top two much less frequently. A Friedman test confirmed that the visualization condition had statistically significant effect on preference rankings, $Q = 29.75, p < 0.001$ with a moderate-to-strong effect size (Kendall's $W = 0.41$). Post-hoc pairwise comparisons using Bonferroni corrections showed that the *Baseline* condition was ranked significantly lower than the *ARSelfie*, *Mannequin* and *BarChart* conditions. No other pairwise comparisons were statistically significant.

7 Discussion and Design Implications

Our findings show that spatial placement of feedback strongly shapes workload, user experience and performance in facial exercises. We interpret these results to derive strategies for designing future facial exercise systems.

7.1 Embedded Feedback for the Face?

Our quantitative and qualitative results indicate that the benefits of embedded feedback, previously demonstrated for whole-body movements, extend to facial movements as well. NASA-TLX and ECL results indicate that the spatial alignment of visual cues with the region of movement in the embedded conditions (*ARSelfie* and *Mannequin*) enabled users to directly associate feedback with the target facial muscles. It is plausible that this would minimize the need for spatial translation [31] between the feedback and the targets, resulting in lower task and cognitive load compared to the *BarChart* and *Baseline* conditions (**RQ2**). In contrast, the *BarChart* condition was cognitively demanding as it required users to mentally translate feedback presented in a peripheral location onto their face to make adjustments. This spatial separation likely increased *cognitive distance* [27] between the feedback and its referent, increasing the chances of the *split-attention* effect [14]. The significantly higher score on the NASA-TLX temporal demand sub-scale for the *BarChart* condition, compared to the *ARSelfie*, further suggests that attention shifts between the face and feedback combined with the spatial translation process contributed to participants feeling more rushed. Although participants were more accurate in activating their muscles while using the *BarChart* condition, the performance gain likely came at the cost of learning and cognitive effort. The higher perspicuity and efficiency scores for the embedded visualizations further indicate that spatially congruent feedback was easier to integrate. This aligns with the *spatial continuity principle* [36].

Designers of facial exercise feedback systems should leverage spatially congruent visualizations that align directly with the target muscles as they reduce extraneous cognitive and task load imposed on the user (**RQ2**). While situated feedback modalities support accurate movements (**RQ1**), designers must be mindful of the cognitive demands associated with the attention shifts and spatial translation. Broadly, these findings indicate that embedded feedback design principles, established for whole-body or upper-limb training can be extended to facial exercises.

7.2 Selfie Tradeoffs

The embedded conditions were generally preferred to the *BarChart* and *Baseline* conditions **(RQ3)** with the *ARSelfie* condition rated highest in terms of user experience and the most preferred overall, indicating the effectiveness of spatially aligned feedback for facial movement tasks. However, qualitative feedback revealed important tradeoffs for designing embedded feedback. Even though the participants valued the spatial benefits of the embedded conditions, their opinions on the selfie-view varied considerably. Some found visual overlays on their face interactive and intuitive, but for others seeing their face was distracting and uncomfortable.

These contrasting responses align with research on self-focused attention, which suggests that individuals are hardwired to attend to their faces and this can override top-down attentional control and lead to cognitive exhaustion [8, 57]. This could be compounded by self-evaluation, mirror anxiety and "Zoom fatigue," where sustained focus on one's own image can trigger negative self-perception and deplete attentional resources [3, 10]. In this state, a user's attention may be captured by their own perceived flaws, rendering them effectively blind to the AR feedback itself [30]. These concerns might have partially offset the benefits of spatial alignment in some participants.

The *Mannequin* condition mitigated this by providing a less personal proxy. While this approach did reduce self-consciousness for some, others described the avatar as "unnatural" or "creepy", indicating sensitivity to uncanny valley effects [37].

Our findings surface a core design trade-off. While overlaying feedback on a face, whether the user's own or an avatar's, might provide superior spatial guidance, its effectiveness can be enhanced by accommodating individual differences in self-representation comfort.

We recommend that embedded facial feedback systems incorporate flexible self-representation options to maximize the spatial benefits while minimizing psychological barriers. This could include progressive abstraction levels (from stylized avatars to photorealistic), customizable avatar features to reduce uncanny valley effects, using proxies of favorite fictional characters, or hybrid approaches that combine abstract overlays with minimal facial landmarks. This could allow users to fully benefit from the advantages of embedded feedback.

7.3 Perceptual Clarity

While our results indicate that the benefits of embedded feedback extend to the face, their effectiveness is highly dependent on the perceptual clarity of the cues.

First, participants struggled to interpret the magnitude of corrective adjustment conveyed through opacity modulation. This encoding limitation created a preference-performance trade-off, while users preferred embedded conditions for its spatial advantages, they activated their muscles more accurately leveraging the explicit magnitude representation in the *BarChart* condition. Second, the AR overlay itself introduced some visual clutter. Some participants found that the semi-transparent face filters obscured the view of the facial muscles they were trying to control. This visual interference can be detrimental for learning, as clear visual feedback is critical for developing accurate motor control especially during the skill acquisition stage [41].

Therefore, we recommend designers combine spatially embedded feedback with explicit magnitude indicators. While spatial cues

are effective for showing where the users should focus, they should be supplemented with perceptually effective cues such as proportional fill of the muscle area or outlines or numerical percentages in text to clearly communicate how much adjustment is needed to reach the target. These filters must be designed carefully to provide feedback without obscuring the user's self view.

7.4 Feedback Geometry

The *BarChart* condition used 2D bars with consistent shape and size for all muscles, whereas the embedded conditions used overlays that followed the natural, 3D contours of individual facial muscles. This difference could have influenced how participants perceived and interpreted the feedback. The bar fills potentially benefited from their uniform geometry as changes were visually consistent across muscles and hence, easier to compare. But it required frequent attentional shifts.

In contrast, the overlays provided spatially aligned feedback that made it easier for participants to connect it with the muscles they were exercising but these overlays varied in shape and extent. The changes in larger muscles (e.g., forehead) could have been more conspicuous while the changes in small or thin muscle overlays were subtle and harder to detect. This uneven salience of feedback across muscles presents a challenge unique to facial embedded feedback, while spatial alignment reduces effort and enhances user experience, geometric variability of the overlays can create perceptual imbalances.

When using embedded feedback in facial exercise systems designer should ensure that they normalize perceptual salience by adjusting visual parameters such as brightness or contrast to amplify feedback on smaller or less conspicuous muscles. They can also use supplementary indicators like subtle outlines or glyphs to provide consistent salience across muscles.

8 Limitations and Future Work

Our work indicates that spatially aligned, embedded visualizations can effectively support facial motor training by reducing cognitive load and enhancing user experience compared to situated visualizations. However, several limitations qualify our findings and point to future directions.

First, the study was conducted in a controlled laboratory setting lasting 45–60 minutes and focused on three exercises involving 11 distinct facial muscles. While this scope allowed focused comparison across visualization conditions, it leaves open questions about the long-term effectiveness of embedded feedback in supporting the learning and retraining of facial movements. Future work should assess learning gains associated with each visualization condition through a between-subjects study and explore longitudinal use, broader exercise repertoires, and more ecologically valid settings to assess sustained outcomes.

Second, our evaluation focused exclusively on visual feedback. Even within a short session, some participants reported discomfort when repeatedly viewing their own face in the *ARSelfie* condition. This discomfort may be amplified in clinical or rehabilitation contexts, potentially discouraging engagement. Future systems should examine avatar-based or proxy feedback strategies in greater depth, and investigate non-visual modalities such as auditory or haptic cues—particularly in contexts where repeated face viewing is not viable or desirable.

Third, our feedback pipeline uses the ARCore face mesh with a manual mapping of vertices to underlying musculature and heuristics to infer muscle activation states from detected facial movements, both validated by a facial anatomy expert. While this mapping provides anatomical grounding, it remains an indirect proxy for muscle activation and does not capture physiological signals such as recruitment strength or neuromuscular coordination. Before real-world deployment for facial muscle training applications, this ARCore-based muscle activation detection system and its underlying heuristics should be validated against ground-truth physiological measurements such as EMG to ensure accuracy and reliability. Future work could also use sophisticated algorithms such as [34] to predict muscle activation directly from image data, integrate complementary sensing modalities (e.g., EMG) or incorporate biomechanical models of facial anatomy to enable more precise and potentially diagnostic feedback.

Finally, our participant sample was relatively small and homogeneous in age, which limits generalizability. Future studies should recruit larger and more diverse populations, including clinical groups and individuals with varied expressive norms or cultural backgrounds, to better assess the applicability and inclusivity of our design insights.

9 Conclusion

In this work, we investigated how the spatial placement of visual feedback affects performance, cognitive load, and user experience in facial motor training. We compared three visualization techniques of situated (BarChart), proxy-embedded (Mannequin), and embedded (ARSelfie) and a Baseline condition where participants saw their face in the mobile device's front-facing camera but received no additional feedback on facial muscle activation, in a within-subjects study (N=24). Our results indicate that spatially aligned feedback reduces cognitive effort, improves user experience, and is generally preferred over situated alternatives. While situated feedback can support greater accuracy, it imposes higher cognitive demand due to attention shifts and the need for spatial translation. Based on these findings, we provide design implications for future AR-based facial feedback systems, with implications for rehabilitation, performance training, and skill acquisition.

Acknowledgments

We thank Shweta Dharmatti for helping us with the facial landmarks to muscles mapping, Arthur Caetano for his input on the design of the visualization conditions and members of the UCSB Four Eyes Lab for feedback on the manuscript draft. We thank the U.S. National Science Foundation for supporting this research through the Early CAREER Award 2023 no. 2240133.

References

[1] Mohammed F. Alhamid, Mohamad Eid, and Abdulmotaleb El Saddik. 2012. A multi-modal intelligent system for biofeedback interactions. In *2012 IEEE International Symposium on Medical Measurements and Applications Proceedings*. 1–5. doi:10.1109/MeMeA.2012.6226653

[2] Fraser Anderson, Tovi Grossman, Justin Matejka, and George Fitzmaurice. 2013. YouMove: enhancing movement training with an augmented reality mirror. In

Proceedings of the 26th annual ACM symposium on User interface software and technology. 311–320.

[3] Jeremy N Bailenson. 2021. Nonverbal overload: A theoretical argument for the causes of Zoom fatigue. (2021). doi:10.1037/tmb0000030

[4] Giuliana Barrios Dell'Olio and Misha Sra. 2021. FaraPy: An Augmented Reality Feedback System for Facial Paralysis using Action Unit Intensity Estimation. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. ACM, Virtual Event USA, 1027–1038. doi:10.1145/3472749.3474803

[5] Matthew Berry and Steven Brown. 2022. The dynamic mask: Facial correlates of character portrayal in professional actors. *Quarterly Journal of Experimental Psychology* 75, 5 (May 2022), 936–953. doi:10.1177/17470218211047935 Publisher: SAGE Publications.

[6] Susana Bloch. 1993. Alba Emoting: A psychophysiological technique to help actors create and control real emotions. *Theatre Topics* 3, 2 (1993), 121–138.

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[8] Serge Brédart, Marie Delchambre, and Steven Laureys. 2006. Short article: one's own face is hard to ignore. *Quarterly journal of experimental psychology* 59, 1 (2006), 46–52.

[9] Xueyan Cai, Yingjing Xu, Zihong Zhou, Mengru Xue, Zhengke Li, Chentian Weng, Wei Luo, Cheng Yao, Bo Lin, and Jianwei Yin. 2025. "Break the Mask Barrier": An AU-based Rehabilitation Training System for Parkinson's Hypomimia. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–9. doi:10.1145/3706599.3719706

[10] Alecia Carolli. 2023. *Zoom Usage and Cognitive Load.* Master's thesis. York University.

[11] Polona Caserman, Shule Liu, and Stefan Göbel. 2022. Full-Body Motion Recognition in Immersive- Virtual-Reality-Based Exergame. *IEEE Transactions on Games* 14, 2 (2022), 243–252. doi:10.1109/TG.2021.3064749

[12] Luigi Cattaneo and Giovanni Pavesi. 2014. The facial motor system. *Neuroscience & Biobehavioral Reviews* 38 (Jan. 2014), 135–159. doi:10.1016/j.neubiorev.2013.11.002

[13] Luigi Ceccaroni and Xavier Verdaguer. 2004. Magical Mirror: multimedia, interactive services in home automation. In *Proceedings of the workshop on environments for personalized information access*. 10–21.

[14] Paul Chandler and John Sweller. 1992. The split-attention effect as a factor in the design of instruction. *British Journal of Educational Psychology* 62, 2 (1992), 233–246.

[15] Lei Chen, Chee Wee Leong, Gary Feng, Chong Min Lee, and Swapna Somasundaran. 2015. Utilizing multimodal cues to automatically evaluate public speaking performance. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. 394–400. doi:10.1109/ACII.2015.7344601 ISSN: 2156-8111.

[16] Alana Elza Fontes Da Gama, Thiago Menezes Chaves, Lucas Silva Figueiredo, Adriana Baltar, Ma Meng, Nassir Navab, Veronica Teichrieb, and Pascal Fallavollita. 2016. MirrARbilitation: A clinically-related gesture recognition interactive tool for an AR rehabilitation system. *Computer methods and programs in biomedicine* 135 (2016), 105–114.

[17] Florian Diller, Nico Henkel, Gerik Scheuermann, and Alexander Wiebel. 2025. SkillAR: omnipresent in-situ feedback for motor skill training using AR. *Virtual Reality* 29, 1 (Feb. 2025), 33. doi:10.1007/s10055-025-01108-1

[18] Paul Ekman and Erika L Rosenberg. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

[19] Paul M Fitts and Michael I Posner. 1967. Human performance. (1967).

[20] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction* 35, 6 (2019), 456–467.

[21] Miharu Fuyuno, Rinko Komiya, and Takeshi Saitoh. 2018. Multimodal analysis of public speaking performance by EFL learners: Applying deep learning to understanding how successful speakers use facial movement. *The Asian Journal of Applied Linguistics* 5, 1 (April 2018), 117–129. https://caes.hku.hk/ajal/index.php/ajal/article/view/508 Number: 1.

[22] Gissela M. Guanoluisa, Jimmy A. Pilatasig, and Víctor H. Andaluz. 2019. GY MEDIC: Analysis and Rehabilitation System for Patients with Facial Paralysis. In *Integrated Uncertainty in Knowledge Modelling and Decision Making*, Hirosato Seki, Canh Hao Nguyen, Van-Nam Huynh, and Masahiro Inuiguchi (Eds.). Springer International Publishing, Cham, 63–75. doi:10.1007/978-3-030-14815-7_6

[23] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*, Peter A. Hancock and Najmedin Meshkati (Eds.). Human Mental Workload, Vol. 52. North-Holland, 139–183. doi:10.1016/S0166-4115(08)62386-9

[24] Koki Hasebe, Tsuyoshi Kojima, Yusuke Okanoue, Ryohei Yuki, Hirotaka Yamamoto, Shuya Otsuki, Shintaro Fujimura, and Ryusuke Hori. 2024. Novel evaluation method for facial nerve palsy using 3D facial recognition system in iPhone. *Auris Nasus Larynx* 51, 3 (June 2024), 460–464. doi:10.1016/j.anl.2024.02.003

[25] Pedro Henriquez, Bogdan J. Matuszewski, Yasmina Andreu-Cabedo, Luca Bastiani, Sara Colantonio, Giuseppe Coppini, Mario D'Acunto, Riccardo Favilla, Danila Germanese, Daniela Giorgi, Paolo Marraccini, Massimo Martinelli, Maria-Aurora Morales, Maria Antonietta Pascali, Marco Righi, Ovidio Salvetti, Marcus Larsson, Tomas Strömberg, Lise Randeberg, Asgeir Bjorgan, Giorgos Giannakakis, Matthew Pediaditis, Franco Chiarugi, Eirini Christinaki, Kostas Marias, and Manolis Tsiknakis. 2017. Mirror Mirror on the Wall... An Unobtrusive Intelligent Multisensory Mirror for Well-Being Status Self-Assessment and Visualization. *IEEE Transactions on Multimedia* 19, 7 (July 2017), 1467–1481. doi:10.1109/TMM.2017.2666545

[26] Atsuki Ikeda, Dong-Hyun Hwang, Hideki Koike, Gerd Bruder, Shunsuke Yoshimoto, and Sue Cobb. 2018. AR based Self-sports Learning System using Decayed Dynamic TimeWarping Algorithm.. In *ICAT-EGVE*. 171–174.

[27] SeungJun Kim and Anind K. Dey. 2009. Simulated augmented reality windshield display as a cognitive mapping aid for elder driver navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 133–142. doi:10.1145/1518701.1518724

[28] Melina Klepsch, Florian Schmitz, and Tina Seufert. 2017. Development and Validation of Two Instruments Measuring Intrinsic, Extraneous, and Germane Cognitive Load. *Frontiers in Psychology* 8 (Nov. 2017). doi:10.3389/fpsyg.2017.01997 Publisher: Frontiers.

[29] Felix Kosmalla, Florian Daiber, Frederik Wiehr, and Antonio Krüger. 2017. Climbvis: Investigating in-situ visualizations for understanding climbing movements by demonstration. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces*. 270–279.

[30] Carina Kreitz, Philip Furley, Daniel Memmert, and Daniel J Simons. 2015. Inattentional blindness and individual differences in cognitive abilities. *PLOS one* 10, 8 (2015), e0134675.

[31] Axel Larsen and Claus Bundesen. 1998. Effects of spatial separation in visual pattern matching: Evidence on the role of mental translation. *Journal of Experimental Psychology: Human Perception and Performance* 24, 3 (1998), 719–731. doi:10.1037/0096-1523.24.3.719 Place: US Publisher: American Psychological Association.

[32] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and Evaluation of a User Experience Questionnaire. In *HCI and Usability for Education and Work*, Andreas Holzinger (Ed.). Springer, Berlin, Heidelberg, 63–76. doi:10.1007/978-3-540-89350-9_6

[33] Benjamin Lee, Michael Sedlmair, and Dieter Schmalstieg. 2023. Design Patterns for Situated Visualization in Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics* (2023), 1–12. doi:10.1109/TVCG.2023.3327398

[34] Ruofan Liu, Yichen Peng, Takanori Oku, Chen-Chieh Liao, Erwin Wu, Shinichi Furuya, and Hideki Koike. [n. d.]. From Pose to Muscle: Multimodal Learning for Piano Hand Muscle Electromyography. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[35] Tania Marur, Yakup Tuna, and Selman Demirci. 2014. Facial anatomy. *Clinics in Dermatology* 32, 1 (Jan. 2014), 14–23. doi:10.1016/j.clindermatol.2013.05.022

[36] Richard E Mayer and Logan Fiorella. 2014. 12 principles for reducing extraneous processing in multimedia learning: Coherence, signaling, redundancy, spatial contiguity, and temporal contiguity principles. *The Cambridge handbook of multimedia learning* 279 (2014), 279–315.

[37] Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & automation magazine* 19, 2 (2012), 98–100.

[38] Mark Nestor, Glynis Ablon, and Andy Pickett. 2017. Key Parameters for the Use of AbobotulinumtoxinA in Aesthetics: Onset and Duration. *Aesthetic Surgery Journal* 37, suppl_1 (May 2017), S20–S31. doi:10.1093/asj/sjw282

[39] Donald A Norman. 1986. Cognitive engineering. In *User centered system design*. CRC Press, 31–62.

[40] Sebastian Rigling, Xingyao Yu, and Michael Sedlmair. 2023. "In Your Face!": Visualizing Fitness Tracker Data in Augmented Reality. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–7. doi:10.1145/3544549.3585912

[41] Christelle Robin, Lucette Toussaint, Yannick Blandin, and Annie Vinter. 2004. Sensory integration in the learning of aiming toward "self-defined" targets. *Research Quarterly for Exercise and Sport* 75, 4 (2004), 381–387.

[42] Richard A Schmidt and Craig A Wrisberg. 2008. *Motor learning and performance: A situation-based learning approach*. Human kinetics.

[43] Martin Schrepp, Andreas Hinderks, and Jorg Thomaschewski. 2017. Construction of a benchmark for the user experience questionnaire (UEQ). *International Journal of Interactive Multimedia and Artificial Intelligence* 4, 4 (2017), 40–44. doi:10.9781/ijimai.2017.445

[44] Yoones A Sekhavat and Mohammad S Namani. 2018. Projection-based ar: Effective visual feedback in gait rehabilitation. *IEEE Transactions on Human-Machine Systems* 48, 6 (2018), 626–636.

[45] Todd K Shackelford and Randy J Larsen. 1997. Facial asymmetry as an indicator of psychological, emotional, and physiological distress. *Journal of personality and social psychology* 72, 2 (1997), 456.

[46] Lun Shu, Victor R. Barradas, Zixuan Qin, and Yasuharu Koike. 2025. Facial expression recognition through muscle synergies and estimation of facial keypoint displacements through a skin-musculoskeletal model using facial sEMG signals. *Frontiers in Bioengineering and Biotechnology* 13 (Feb. 2025). doi:10.3389/fbioe.2025.1490919 Publisher: Frontiers.

[47] Roland Sigrist, Georg Rauter, Robert Riener, and Peter Wolf. 2013. Augmented visual, auditory, haptic, and multimodal feedback in motor learning: a review. *Psychonomic bulletin & review* 20, 1 (2013), 21–53.

[48] Rajinder Sodhi, Hrvoje Benko, and Andrew Wilson. 2012. LightGuide: projected visualizations for hand movement guidance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 179–188. doi:10.1145/2207676.2207702

[49] Maurício Sousa, João Vieira, Daniel Medeiros, Artur Arsenio, and Joaquim Jorge. 2016. SleeveAR: Augmented Reality for Rehabilitation using Realtime Feedback. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. ACM, Sonoma California USA, 175–185. doi:10.1145/2856767.2856773

[50] Hariharan Subramonyam. 2015. SIGCHI: Magic Mirror - Embodied Interactions for the Quantified Self. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, Seoul Republic of Korea, 1699–1704. doi:10.1145/2702613.2732884

[51] John Sweller. 1994. Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction* 4, 4 (1994), 295–312.

[52] Richard Tang, Xing-Dong Yang, Scott Bateman, Joaquim Jorge, and Anthony Tang. 2015. Physio@ Home: Exploring visual guidance and feedback techniques for physiotherapy exercises. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 4123–4132.

[53] Milka Trajkova and Francesco Cafaro. 2018. Takes Tutu to ballet: designing visual and verbal feedback for augmented mirrors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–30.

[54] Sean Michael White. 2009. *Interaction and presentation techniques for situated visualization*. Columbia University.

[55] Wesley Willett, Yvonne Jansen, and Pierre Dragicevic. 2016. Embedded data representations. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 461–470.

[56] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 143–146.

[57] Michał J Wójcik, Maria M Nowicka, Ilona Kotlewska, and Anna Nowicka. 2018. Self-face captures, holds, and biases attention. *Frontiers in psychology* 8 (2018), 2371.

[58] Fan Xu, Xian-wei Zou, Li-qiong Yang, Shi-cong Mo, Quan-hao Guo, Jing Zhang, Xiechuan Weng, and Guo-gang Xing. 2022. Facial muscle movements in patients with Parkinson's disease undergoing phonation tests. *Frontiers in Neurology* 13 (Oct. 2022). doi:10.3389/fneur.2022.1018362 Publisher: Frontiers.

[59] Xingyao Yu, Benjamin Lee, and Michael Sedlmair. 2024. Design Space of Visual Feedforward And Corrective Feedback in XR-Based Motion Guidance Systems. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–15. doi:10.1145/3613904.3642143

[60] Xingyao Yu, David Rosin, Johannes Kässinger, Benjamin Lee, Frank Dürr, Christian Becker, Oliver Röhrle, and Michael Sedlmair. 2024. PerSiVal: On-Body AR Visualization of Biomechanical Arm Simulations. *IEEE Computer Graphics and Applications* 44, 6 (Nov. 2024), 24–38. doi:10.1109/MCG.2024.3494598

[61] Junyi Zhu, Yuxuan Lei, Aashini Shah, Gila Schein, Hamid Ghaednia, Joseph Schwab, Casper Harteveld, and Stefanie Mueller. 2022. MuscleRehab: Improving unsupervised physical rehabilitation by monitoring and visualizing muscle engagement. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–14.