The Motion is the Message: Evaluating Motion Tracking Quality for VR Avatars

Fu-Chia Yang, Harrison Jesse Smith, Christos Mousas, and Michael Neff

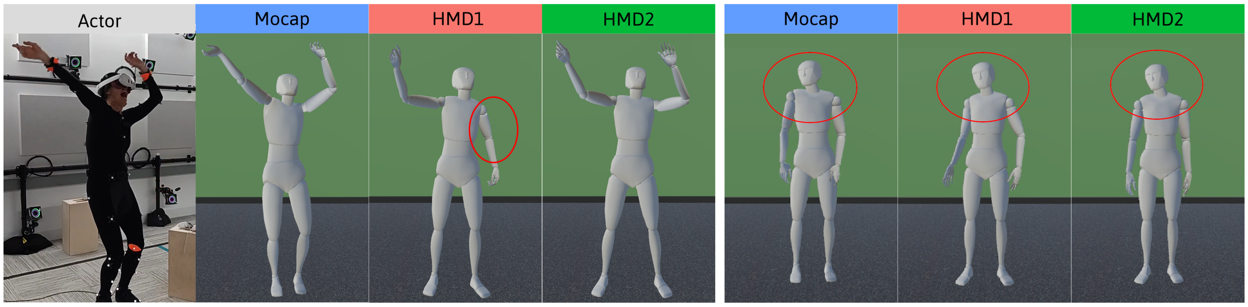


Fig. 1: Left: Motion is simultaneously recorded using marker-based motion capture and head-mounted display (HMD) tracking, allowing us to perceptually evaluate any differences in the social information conveyed. Right: A subtle error in coronal posture tracking that led to participants failing to recognize a happy motion.

Abstract—Motion tracking to project users into embodied virtual reality (VR) as avatars is an essential application of real-time computer graphics. Most current embodied VR systems rely on head-mounted displays (HMDs) to estimate user pose, as headset sensors can track the head and hands, thereby reconstructing the full body without the need for external hardware. However, measuring the quality of motion reconstruction algorithms from HMD-based tracking, particularly those intended for use in social settings, remains challenging due to the complex interaction between motion and perceived social signals. This paper compares two industrial tracking reconstruction solutions, called HMD1 (i.e., a basic HMD-based method that uses head tracking and hand positions estimated from HMD cameras) and HMD2 (i.e., an advanced HMD-based method with additional onboard camera streams), that estimate user motion using only an HMD against ground-truth motion capture (MoCap) data. It advocates for a social signal-based analysis that views motion as a communication medium and employs user observations to measure whether viewers successfully perceive the information encoded in motion. Across 156 socially expressive clips, Social Signal ratings were more effective than generic measures at revealing differences between the HMD methods. HMD2 preserved social meaning more accurately than HMD1, with fewer significant deviations from MoCap, while both HMD methods were frequently rated less natural than MoCap. A qualitative review localized recurrent failure modes, such as arm swivel/shoulder errors, posture reconstruction issues, and floating/stance artifacts, which help explain the misreading of social signals. We release a dashboard scorecard, motion capture data, and a benchmark protocol to enable consistent motion evaluation. More generally, this work advocates for an underexplored approach to motion evaluation that focuses on assessing the semantics of motion to determine quality. As reliance on generative artificial intelligence (AI) increases, it is essential to standardize evaluation to preserve the authenticity of the social signals conveyed. The developed dataset and the evaluation framework are provided on our project's website: https://github.com/facebookresearch/MotionIsTheMessageDataset.

Index Terms—Virtual reality, motion tracking, evaluation, naturalness.

1 Introduction

Embodied virtual reality (VR) [45, 47] enables users to interact in virtual environments by having their avatars controlled through their physical movements, which are captured and mirrored in real time. Early systems, especially those using head-mounted displays (HMDs), often relied on limited "three-point tracking" (3pt), in which only the head and two hand controllers were tracked and the lower-body pose was inferred from this sparse data [39, 61]. Recent HMDs integrate additional onboard cameras that improve tracking fidelity, enabling more accurate reconstruction of full-body motion and fingers compared to 3pt tracking, while lower-body pose estimation remains limited. Crucially, these advances allow tracking with only an HMD, removing the need for handheld controllers or external devices.

However, beyond technical progress, improved tracking has important consequences for social interaction in embodied VR, where nonverbal cues (e.g., emotional expression, attentiveness, dominance) play a central role in effective communication. Avatars must accurately reproduce these social signals, since failure to do so risks misunderstandings and a diminished sense of social presence.

Evaluating animation quality in social contexts is, therefore, a nontrivial task. For physically demanding movements (e.g., tumbling, parkour), fidelity can be assessed with metrics that measure adherence to physical constraints, such as the conservation of angular momentum. However, when motion conveys social meaning rather than just physical precision, such metrics become insufficient. In these cases, the evaluation must determine whether the intended social messages embedded in the motion are preserved and interpretable to observers.

We address this methodological gap (i.e., the lack of evaluation approaches that directly measure the preservation of social meaning) by proposing an evaluation framework grounded in perceptual judgments of social signal communication. Specifically, we compared three distinct motion-tracking solutions: (1) HMD1, which uses the headset to

- *Fu-Chia Yang and Christos Mousas are with Purdue University. E-mail: {yang1684 | cmousas}@purdue.edu.*
- *Harrison Jesse Smith is with Meta Inc. E-mail: hjessmith@gmail.com.*
- *Michael Neff is with Meta Inc. and the University of California, Davis. E-mail: mpneff@ucdavis.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

track the head pose and HMD cameras for hand and finger tracking, from which upper-body poses are reconstructed, with lower-body poses reconstructed at limited fidelity; (2) HMD2, an enhanced approach that leverages additional onboard camera streams to substantially improve upper-body pose accuracy relative to HMD1, resulting in an improved but still limited lower-body reconstruction; and (3) MoCap, a high-quality marker-based optical motion capture system that serves as a ground-truth baseline. Both HMD systems are commercially released solutions, highlighting the practical relevance of our analysis.

To support this comparison, we captured a comprehensive set of socially expressive motions performed by professional actors. Motions were organized into: (1) Signal scales, bipolar dimensions designed to convey specific social states (e.g., *Energetic–Tired*); (2) Posture scales, arm posture and posture range; and (3) Emotion scales. We identified body components critical for conveying social signals (e.g., posture, arms, fingers, shoulders, and head), and designed the scales to ensure that fidelity across all upper-body regions would be exercised by the motion set. All motions were recorded simultaneously with both HMD streams and the optical motion capture system, allowing direct comparison of tracking fidelity and social signal preservation.

We conducted a perceptual evaluation study in VR in which participants viewed avatar animations generated from each tracking solution. Conducting the study in VR allows participants to perceive motions under a spatially immersive environment, where depth cues, scale, and co-presence are naturally supported. These factors influence motion perception and social interpretation, and are limited in 2D displays. Participants rated (a) the perceived accuracy with which social signals were communicated and (b) overall naturalness of the motion. These ratings jointly assess the fidelity of each tracking system and its effectiveness at preserving and communicating realistic social interaction. Through the first of these criteria, we explore an *encoding-decoding* approach to evaluation, in which motion is viewed as encoding a particular signal and success is judged based on whether that signal can be successfully decoded by observers.

We situate our study within a broader agenda: *in animation settings where social communication is paramount, visual fidelity alone is insufficient*. With the rapid adoption of generative artificial intelligence (GenAI) for motion synthesis, it will become increasingly important to develop evaluation approaches that assess whether subtle, user-intended social signals are preserved. Rather than evaluating GenAI synthetic motions directly, our work takes a first step toward this broader challenge by focusing on evaluating whether current HMD-based tracking reconstructions can preserve socially meaningful motion. To this end, we formulated the following research questions:

- **RQ1:** Do HMD1 and HMD2 accurately convey the social information contained in the motion, relative to MoCap as ground truth?
- **RQ2:** Are HMD1 and HMD2 natural, relative to MoCap as ground truth?
- **RQ3:** Do Social Signal and Naturalness ratings provide similar guidance on tracking performance?
- **RQ4:** Do the developed scales provide insight into the motion reconstruction algorithms when HMD1 and HMD2 produce weaker ratings, relative to MoCap as ground truth?

In summary, this paper makes the following three key contributions, which provide a generalizable framework for assessing whether motion retains its communicative intent in virtual environments: (1) an evaluation of a critical motion-tracking application, demonstrating that limitations in tracking reconstruction quality can impede the perception of social signals; (2) a perceptually grounded testing methodology that focuses on how well social meaning is preserved in motion; and (3) a curated dataset and benchmark to support future research into socially expressive motion, enabling direct comparisons across future tracking, reconstruction, and synthesis approaches.

2 Related Work

2.1 Motion Evaluation

There has been a long history of motion evaluation in animation [16, 19, 20, 41], with a recent survey offered by Rekik et al. [42]. One area that is closely aligned with the social applications targeted here is gesture synthesis evaluation. A review of the field by Wolfert et al. [59] found that most studies are within-group and use subjective evaluations; however, no systematic approach has been adopted for these studies. Quantitative measures included gesture speed, trajectory, jerk, and the Frechet Gesture Distance (FGD). Qualitative measures included perceived naturalness, appropriateness of gesture timing, and the Godspeed questionnaire [3], which quantifies the human likeness, animacy, likability, and perceived intelligence, as well as the impact of gesture on recall and comprehension. They argue for including a ground truth condition. The GENEA Challenge, part of the GENEA Initiative,[1] is a notable effort to standardize the evaluation of gesture synthesis systems [28, 29]. It provides researchers with a common dataset, model, and testing environment, and then conducts crowdsourced evaluations of the output they produce, assessing the human-likeness (i.e., naturalness) and appropriateness of the gesture for the speech.

2.2 Automatic and Subjective Evaluations

It would be desirable to have a quantitative measure of motion quality, as this would enable automatic evaluation, where motion quality can be assessed algorithmically without relying on human raters, facilitating system comparison, and potential integration into learning frameworks. Unfortunately, no quantitative measures are currently well justified in this area. Voas et al. [52] argued that existing evaluation metrics for assessing the accuracy of skeleton-based human motion, generated from natural language descriptions, exhibit low correlation with human judgments at the sample level, although some metrics perform better when averaged across models. Kucherenko [29] found that none of the metrics, including average jerk, average acceleration, distance between gesture speed (i.e., absolute velocity) histograms, or Canonical Correlational Analysis (CCA), provided a ranking of gesture synthesis algorithms that significantly correlated with human ratings. The Frechet Inception Distance (FID) was introduced for image synthesis using generative adversarial networks (GANs) [18]. It compares the distribution of synthesized output against the ground truth. For the gesture-related measure, the FGD was introduced by Yoon et al. [62]. Analysis of the 2022 GENEA Challenge found positive evidence for the FGD [29]. However, Tseng et al. [51] did not find that FID models were reliable for dance, and Voas et al. [52] also found that FID did not perform well on naturalness. Specifically, Voas et al. [52] developed a machine learning (ML) algorithm called MoBERT, which was trained to perform evaluations based on user ratings of motion naturalness and faithfulness. To our knowledge, automatic evaluation of social signals has not been adequately demonstrated.

Prior work has also explored the design of subjective evaluations. Ferstl et al. [13] compared several different rating scales for subjective evaluation and provided evidence that Naturalness can separate conditions well. Other works have explored different stimulus presentation approaches, including paired simultaneous, sequential, and parallel presentation of many examples [23, 58]. Behavioral measures, like gaze and electroencephalography (EEG), have also been explored [48]. Other work compared free text entry to rating scales [30]. Other issues of presentation, such as viewing perspective [26, 49] and lighting [56] have been examined. People may be less sensitive to motion errors if they are engaged in a task with the character [13]. Many works have focused on error detection [13, 19, 40, 50] in animation data. Recent work has explored errors caused by mismatches between the motion and character model [22, 24, 25, 38, 60]. Other work has compared different model representations, such as differences in gender and ethnicity [21], realistic and cartoony talking heads [27], realistic figures versus robots [14], or full-body versus head and hands representations [17]. Other works have focused on specific tasks in VR, such as

[1] https://genea-workshop.github.io/

avatar-object interaction [6] or manipulation with different inverse kinematics (IK) algorithms [12, 63]. Some works have provided evidence that people interpret motion errors as shifts in personality [13, 38].

2.3 Motion Evaluations in VR

For VR applications, a sense of embodiment in the self-avatar is important and has been extensively studied [11,50,54]. In this work, we focus instead on understanding the social signal presented by a third-person avatar. A closely related study by Adkins et al. [2] examined how finger motion affects comprehension during a charades game. They found that while the presence of hand motion was important for understanding, motion inaccuracies did not significantly impair comprehension. However, jittery hand movements led to discomfort. Despite a broad body of research on how character motion shapes perception, few studies have examined whether social signals, such as personality [36, 37, 46, 57] and emotion [7, 8], are accurately preserved.

Rekik et al. [42] listed nine different terms that have been used as general measures of virtual human animation quality, including “realism", “believability", “naturalness", and “human-likeness". We included Naturalness as one measure in our experiment, but found that it does not adequately capture the issues in motion tracking. Our primary approach, instead, is to focus on social signals that fully engage the upper body, encode these signals carefully, and test whether users can decode them. Some previous work has focused on measuring motion features, which are more aligned with our approach. For example, Wallraven et al. [53] evaluate the intensity and sincerity of facial expressions, while Zibrek et al. [64] consider qualities such as the character’s perceived empathy when evaluating avatar appearance.

Based on prior research and developed evaluation metrics, we designed our study focusing on evaluating social signal accuracy and perceived naturalness for motions that fully engage the upper body in social settings. Our study addresses the current research gap and provides insight into the evaluation of HMD-based motion reconstruction methods.

3 Materials and Methods

3.1 Participants

In total, 62 people between the ages of 18 and 65 participated in the study, with a mean age of 36.6 ($SD = 11.2$). They were recruited from the local community and compensated for their time. Recruitment criteria included proficiency in written and spoken English, not being prone to motion sickness, having no vision-related impairments that would interfere with the VR experience, and the ability to stand for extended periods of time. Thirty-four were female, and 28 were male. In terms of racial background, 23 identified as White/Caucasian, 20 as Asian/Asian American, seven as Black/African/African American, one as Latin/Hispanic, one as Pacific Islander/Native Hawaiian, one as American Indian/Alaska Native, six were mixed race (i.e., four Asian/Caucasian, one African/Caucasian, and one White/Hispanic) and three selected “Other.” In terms of education, 34 had a four-year college degree (e.g., B.A., B.S.), 11 had a Master’s (i.e., M.A., M.S.), PhD, or JD, nine had a two-year college degree (i.e., A.A.), six had some college education, one had a high school diploma, and one had some high school or less education. Most had some experience with VR, with three reporting that they “Have my own VR Hardware,” 32 selecting they used VR “Some times (1-3 times),” 18 selecting “Several times (4+ times),” and nine selecting “No prior experience.”

3.2 Stimuli Preparation

To test whether social signals are accurately maintained by tracking reconstructions, it is necessary to record motions that contain clear social signals. Such signals are a constant part of everyday communication, so one approach to collecting stimuli is to simply record people in interaction. This approach has strengths in terms of ecological validity; however, it is challenging to capture motion that encompasses a wide range of social signals, occurs at varying intensities, and incorporates different aspects of movement when recording spontaneous interactions. Instead, we take a more intentional approach to stimulus collection, as we discuss in this section.

3.2.1 Defining Motion Scales

We defined a set of motion scales. These vary a particular social signal and are performed at different levels between two poles. For example, one scale was from *Tired* to *Energetic*. The chosen scales were selected for two reasons. First, they needed to contain an important signal that supports social interaction. Second, we sought a collection of motions that utilized different parts of the body to thoroughly test the reconstruction algorithms and identify where they fail. We sought motions that relied on posture, shoulder motion, head motion, arm motion, and motion of the character root. A large set of scales was first brainstormed and analyzed to see which portions of the body they exercised. A final group of motions was selected to cover the entire body. This analysis was primarily based on the performing arts literature [35]. For example, it has been reported that the distance the elbow is held from the torso affects the impression of dominance [44], which led to the development of our Arm Posture scale. For each scale, we recorded several takes at each of the designed levels. In the *Energetic* to *Tired* scale, we recorded *Very Tired*, *Tired*, *Neutral*, *Energetic*, and *Very Energetic* levels. This allows us to explore both strong and subtle motions. In addition, since emotion plays a central role in how motion is perceived and interpreted socially, we added the emotion scale. Five emotions were presented using two intensity levels (e.g., *Happy* and *Very Happy*), rather than opposing poles with multiple levels, since bipolar representations can be ambiguous and may overlap with other emotions (e.g., the opposite pole of happy could be interpreted as sad). Table 1 lists all the scales used in the final study, along with the level count and the directions given to the actors. The goal in determining the scales was not to be comprehensive, as that is not practically feasible, but to select a reasonable sample. Our final set included six signal scales, two posture scales, and five emotion scales.

Table 1: The final set of six signal scales, two posture scales, and a sampling of five emotions, along with the corresponding levels count, and prompts given to the actors.

Signal Scale (Levels)	**Prompt**
Attracted – Repulsed (5)	Picture an object in front of you that is either appealing or repulsive.
Dominant – Shy (5)	Greet another person.
Relaxed – Nervous (5)	Ask for an explanation.
Interested – Disinterested (5)	Listen to another.
Energetic – Tired (5)	Plan to get a cup of coffee.
Positive – Negative (5)	React to a present offer.
Posture Scale (Levels)	**Prompt**
Arm Posture (5)	Posture variations involving arm swivel angle (e.g., bring elbow in towards torso or out away from torso).
Posture Range (7)	Posture variations with the flexion of spine and collarbones in the sagittal plane, from collapsed to an overly erect posture.
Emotion Scale (Levels)	**Prompt**
Happiness (2) Anger (2) Sadness (2) Surprise (2) Fear (2)	Perform the emotion at two different intensities.

3.2.2 Stimuli Recording

We hired trained actors to perform these scales in order to obtain high-quality encodings of the desired social signals. The actors were instructed to make the extreme levels, such as *Very Attracted* or *Very Surprised*, still believable for daily life, not cartoonish. The actors were generally given a line of text, simple prompts, and instructed to perform at a particular level. Further direction was given as necessary until satisfactory performances were obtained. We completed recordings with five actors: three females aged 34, 57, and 60, and two males aged 21 and 62. All actors have extensive professional acting experience, most well over a decade. Actors were selected based on submitted audition videos and chosen for their demonstrated ability to perform well-embodied motions.

3.2.3 Apparatus

All recordings were performed on an 8’ × 14’ motion capture stage. Actors wore snug Lycra motion capture suits to facilitate the attachment of motion capture markers to their bodies. They were recorded using an optical marker-based motion capture system, which captured both bodies and hands. The system consisted of 40 Optitrack cameras, 19 Prime^{x} 41 cameras, and 21 Prime^{x} 22 cameras, operated through Motive 3.0.3 software. Actors wore 57 body markers, following the Optitrack motion capture marker configuration. For hand tracking, 38 finger markers, 19 per hand, were used, based on Han et al.’s [15] marker-based hand tracking system. Motions were captured at 120 fps. Audio was recorded with Lavalier microphones, but was not used, as auditory cues could make social intentions easier to identify even when motion reconstruction is imperfect. This high-quality motion capture provided the top-line motion quality for comparison in the study, serving as a reasonable approximation to ground truth.

Actors also wore Meta Quest 3 HMD, which allowed simultaneous capture of optical motion capture and headset-based tracking. Headset-based tracking recorded motion VRS files [34], which consist of time-stamped sensor data that were later processed for motion reconstruction.

3.2.4 Stimuli Processing

The motion capture and HMD recordings were processed separately to obtain skeleton-based motion reconstructions. A skeleton was fitted to the marker positions provided by the motion capture system using the Momentum library [32]. This yields at each frame a pose consisting of the position and orientation of the root, along with the orientation of all remaining joints. Some light cleanup was performed in Autodesk Maya 2024 on the motion capture data to remove artifacts resulting from the solver falling into different local minima, which introduced small jerks in the motion.

Two different tracking sequences, HMD1 and HMD2, were solved from the same HMD recorded VRS files. HMD1 is the baseline full-body reconstruction provided by Meta’s Movement SDK body tracking feature [31]. HMD2 is a full-body reconstruction that includes a higher fidelity mode and IK adjustments. The higher fidelity mode, Inside-Out Body Tracking (IOBT), supports additional body tracking features such as more accurate elbow positions and spine postures [33]. We compared the newer HMD2 tracking solution to the older HMD1 and our MoCap, which serves as our baseline in this study. All motion clips were saved at 24 fps for the online selection study and interpolated to 72 Hz in Unity for the VR study application to match the HMD refresh rate. Any smoothing introduced by this interpolation is a very minor source of error compared with the differences in the three tracking conditions.

3.2.5 Stimuli Selection

Evaluating a large set of motions with multiple clips at each scale level would provide a comprehensive coverage of the space of social message expressions. For instance, a *Disinterested* scale could be conveyed through head turns or backward torso movement. In addition, such variations strongly rely on both the actors’ internal states and their environments. Different choices of encoding by the actors may result in a similar level of social signal intensity in the viewers. However, creating a benchmark dataset that fully samples the entire space is impractical and would contain so many clips as to be onerous for other researchers to use to test their algorithms. Thus, we opt for a single motion at each scale point, leading us to conduct a video stimulus selection study. While this pragmatic design choice precludes our ability to say whether a motion reconstruction algorithm can maintain a social signal in general, it can still provide evidence of where a motion reconstruction algorithm fails. Given the rate of failure of the two state-of-the-art HMD methods we tested in this paper, it is clear that even this limited benchmark will be of value in evaluating reconstruction algorithms for the foreseeable future.

We combined all recorded clips from the five actors and only excluded clips that were outtakes during capture sessions or exhibited poor quality motion reconstruction. Poor quality occurred for clips with finger-solve issues or mesh intersections that required excessive cleanup. For each designed scale, we created videos of all the good-quality clips and presented them separately to 30 participants for each scale in an online study conducted through Amazon Mechanical Turk (MTurk). People saw each video once and were asked to rate it on a 7-point Likert scale with labels based on the particular social signal (e.g., ranging from *Very Tired* to *Very Energetic* in our running example).

Online participants also rated the naturalness of the clip. Naturalness ratings were used as a filter to remove any highly unnatural clips, dropping those that had a rating of less than four out of seven. The remaining clips were ranked based on their average rating on the scale modality (e.g., *Energetic* vs. *Tired*). We then picked the clips that best spanned the ratings range and were evenly spaced. We picked five clips from each of the signal scales and the arm posture scale, seven from the posture range scale, and two from each of the emotion scales, resulting in a total of 52 clips. This ensured that we had the desired mix of more extreme and moderate stimuli. It also took the selection task out of the hands of the experimenters, avoiding any unintended bias. Note that we did not use the actors’ original intention in making these selections. They generally aligned, but there were cases when, say, an actor’s attempt to be *Very Nervous* was selected as the clip to use for *Nervous* based on the ratings. For evaluation purposes, we are interested in whether clips can be consistently decoded by observers across tracking reconstruction algorithms, not any actor’s ability to encode a particular property, so this approach is justified.

3.3 Procedure

Participants attended a single one-hour VR session during which they rated an animated avatar performing 156 distinct motions. The experience was developed in Unity version 2022.3.39f1. Participants wore Meta Quest 3 HMDs, which features a 2064×2208 resolution per eye, a Snapdragon XR2 Gen 2 processor, and 8GB of RAM. At the beginning of the experience, participants were told to stand above a marker on the virtual floor and remain there throughout to ensure everyone had a common viewpoint. After this, they provided demographic information through an input panel in the experience. The overall purpose of the experiment was explained, and then the participants completed a tutorial. The tutorial explained the survey questions, how they were completed using a laser pointer associated with an input controller, and showed a test motion followed by an example rating question for practice.

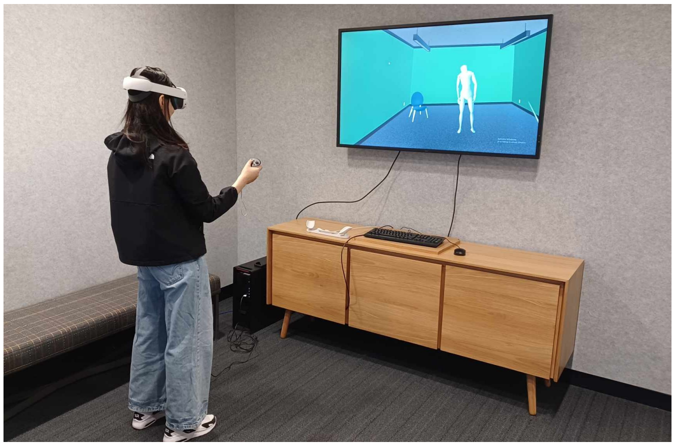
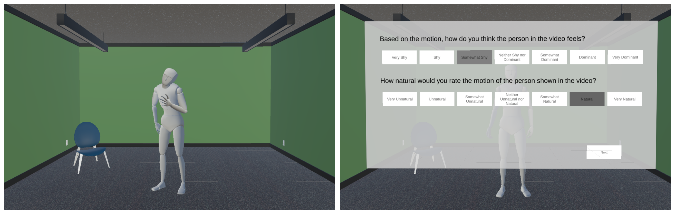


Fig. 2: Participants were immersed in a VR environment to view and provide ratings on the motion stimuli.

During the main experiment, each of the motion scales (see Table 1) was presented in random order. The signal and arm posture range scales consisted of five motions per scale, the posture range scales consisted of seven, and the emotion scales consisted of two per emotion. All motions had been generated at all three quality conditions (i.e., MoCap, HMD1, and HMD2), so participants saw 52 motions (i.e., 30 for signal scales, 12 for posture scales, and 10 for emotion) at three conditions, for a total of 156 motion clips. Clip order was balanced. Participants could click on a "play" button when they were ready to view the clip. The animation was shown on the avatar displayed in Fig. 1, and participants could play it only once. After viewing the clip, participants completed survey questions about the motion. In most cases, they were first asked to rate the motion on its particular scale (e.g. *Energetic* to *Tired*) using a 7-point Likert scale with equally spaced, labeled options (e.g. *Very Tired*, *Tired*, *Somewhat Tired*, *Neither Tired nor Energetic*, *Somewhat Energetic*, *Energetic*, and *Very Energetic*). These were scored from 1 to 7 for analysis. This was followed by the Naturalness rating (i.e., "A motion is natural if it appears like the motion of a real person in this situation. Specifically, you will answer the question: How natural would you rate the motion of the person shown in the video?"), which they could also rate on a 7-point Likert scale.


Based on the motion, how do you think the person in the video feels?

How natural would you rate the motion of the person shown in the video?

Fig. 3: Example of a motion stimulus that was displayed in VR. Participants provided ratings on the motion scale and naturalness through a VR survey after viewing each stimulus clip.

The Emotions scale was a bit different. Participants were first asked to categorize the emotion displayed in the clip (i.e., *Happiness*, *Anger*, *Sadness*, *Surprise*, or *Fear*) and then asked to rate their Confidence in this selection and the clip's Intensity, before rating the Naturalness as with other scales. Each emotion had a more and less intense example.

At the start of each block, participants were given a brief description of the context in which the motion occurred and a preview of the questions they would be asked about the motion. The standard scale question was "Based on the motion, how do you think the person in the clip feels?," followed by a list of Likert labels corresponding to the scale. The provided information is summarized Table 1 in the supplement documentation.

Participants were required to take two-minute breaks after every four scales (i.e., two breaks total). They were invited to sit down during the breaks. The study was approved by the local ethics board.

4 Results

4.1 Analysis Framework

For each of the signal scales and posture scales, participants rated the motion on the scale quality and the motion's naturalness. These can be plotted, as shown in Fig. 4 and Fig. 5. For emotion scales, we plotted the rate of correct identifications, Confidence, Intensity, and Naturalness in Fig. 6.

Our core assumption is that for a tracking method to be successful, it must accurately deliver the message contained in the user's motion, as represented by the motion capture top-line. We therefore considered cases where an HMD tracking result is perceived significantly differently from motion capture as indications of tracking failure. Scales have two factors: the signal (e.g., *Tired* to *Energetic*) and the tracking conditions (i.e., MoCap, HMD1, and HMD2). A traditional two-factor statistical analysis, such as a two-factor ANOVA, would check for variation in each of these factors and then look for significant interactions. However, variance along the scale is designed into the stimuli, so finding it in the results is not informative. We therefore performed our analysis separately at each level of the scale (e.g., we compared HMD1 and HMD2 with MoCap for *Very Disinterested*). This is implemented using Cumulative Link Models [9, 10], a more conservative option that treats the Likert data as ordinal. We used the implementation in the `ordinal` package in `R`. For the emotion identification ratings, we compared whether the user was correct or not, so the data is binomial (i.e., 0 or 1). For this one case, we use a generalized linear mixed effects model with a binomial distribution [4]. Post-hoc effects are tested for by doing pairwise comparisons with expected means using `emmeans` in `R`. Given that we are making multiple comparisons, we need to control for Type I error (i.e., false positives) and do this using False Discovery Rate (FDR) [5]. Bonferroni is a more conservative option, but this choice had minimal impact on the results.

Based on this statistical analysis, we computed two straightforward scores to measure the overall performance of a reconstruction algorithm. The first is the count of the number of times an algorithm produced results statistically different from motion capture; the lower the better, with zero (0) being "perfect" performance. This case count shows how often an algorithm "failed," but it does not measure the size of the failure, which is also relevant. To deepen the analysis, we computed the sum of the effect size for all cases where differences are significant, which indicates how large the difference is between the distribution of responses for the HMD and MoCap conditions. We used Cohen's d as our measure of effect size, which is defined as the difference in distribution means, divided by the overall standard deviation. This provides two scores for the conditions: one captures the number of failure cases, and the other captures the size of the failures.

4.2 Quantitative Results

Table 2 summarizes the scores on each scale. On the count metric, among all 52 motion clips, social signal was perceived significantly differently in 28 cases (i.e., 53.8%) for HMD1 and 21 cases (i.e., 40.4%) for HMD2. The overall effect size score (i.e., sum of Cohen's d [d_T]) is 25.98 for HMD1 and 16.63 for HMD2. For both algorithms, we observed poorer performance compared to motion capture, with significant differences in at least 40% of the cases. Both scores are much better for HMD2. Table 2 also includes the mean effect size for all significant differences. A traditional interpretation of Cohen's d is that a small effect is .20, a medium effect is .50, and a large effect is .80. For HMD1, six scales exceed the large threshold, and three fall between medium and large. For HMD2, three scales exceed the large threshold, and six fall between medium and large. This again suggests that HMD2 performs more closely to motion capture, although both have significant deviations from this gold standard.

It was postulated that the more moderate motion changes might be more difficult for tracking reconstructions to capture than the more extreme ones. There does not appear to be evidence of that in this dataset. Considering the signal scales, the total number of times that HMD1 or HMD2 was worse than the motion capture at the "very" level (i.e., *Very Positive* or *Very Negative*) is 14, with a $d_T = 11.91$. For the more moderate cases, one in from the extreme, the count was 11, with a $d_T = 8.41$. Performance is somewhat worse for the extreme motions. Interestingly, the positive extreme count (i.e., the right end of the charts) was 9 with a $d_T = 8.37$, whereas the negative extreme count was 5 with a $d_T = 3.54$. This suggests tracking was inferior at the extreme right end of the scales.

The right of Table 2 summarizes the results for Naturalness ratings. While for Naturalness, it would be reasonable to perform a two-factor analysis, we elected to use the same analysis as with the scale ratings to allow direct comparison. For Naturalness, higher scores are always better, and the HMD scores were compared against motion capture. For HMD1, the number of clips rated worse than motion capture was 41 (78.8%) with a $d_T = 54.93$. Interestingly, 6 HMD1 clips were rated *more* natural than motion capture with a $d_T = 3.41$. HMD2 clips were never rated more natural than MoCap. The count score for HMD2 is 49 (i.e., worse in 94.2% of cases) with a $d_T = 56.06$. Both HMD reconstructions perform poorly in terms of Naturalness, falling below motion capture in at least 75% of cases. In all significant negative cases, the mean effect size is well above the threshold for a large effect size, indicating a marked drop in Naturalness.

Comparing HMD1 and HMD2, HMD1 seems to fare somewhat better. The d sums are virtually identical (i.e., 54.93 vs. 56.06), but

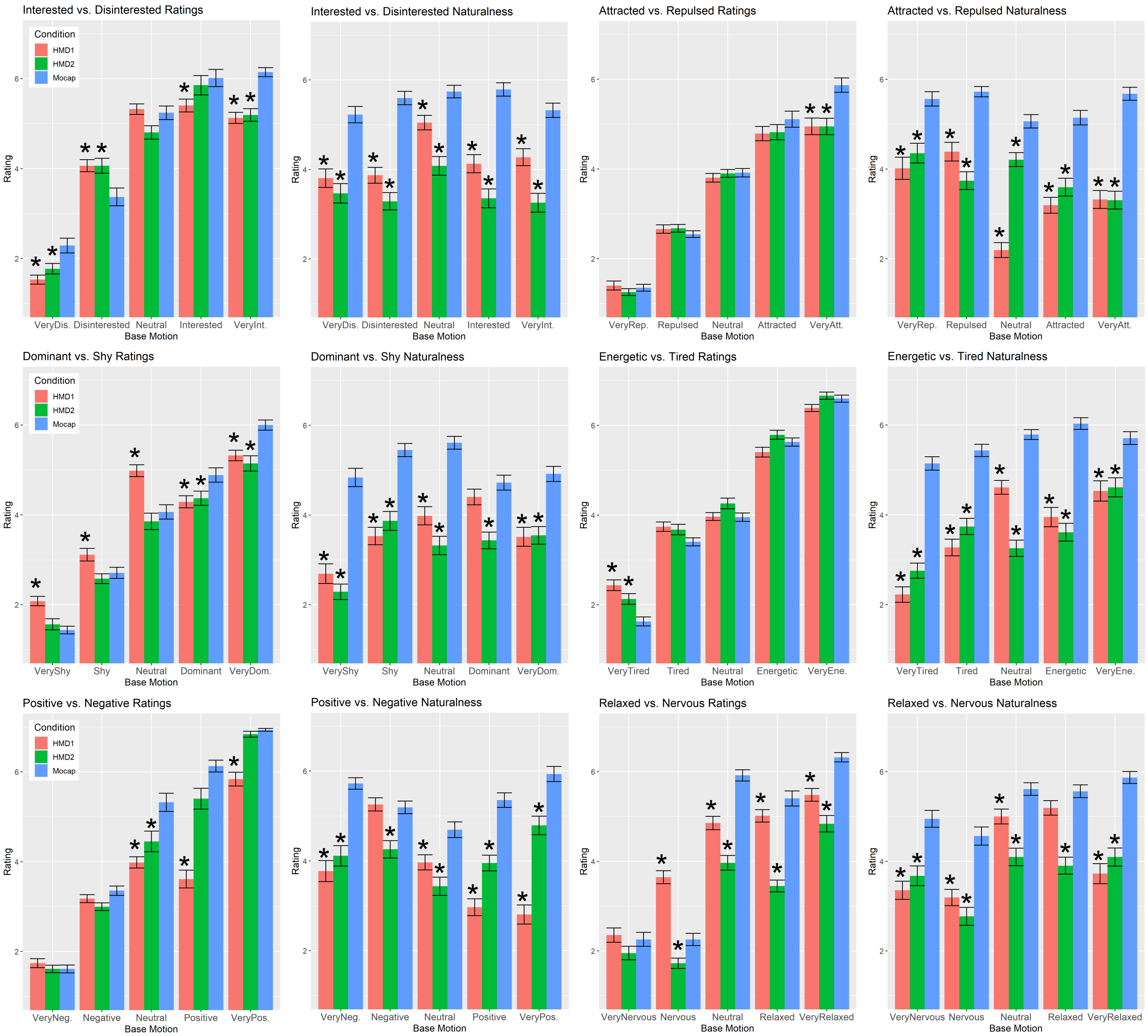


Fig. 4: Social signal and Naturalness ratings for signal scales. * indicates HMD clips that performed significantly differently from MoCap. Error bars show standard error.

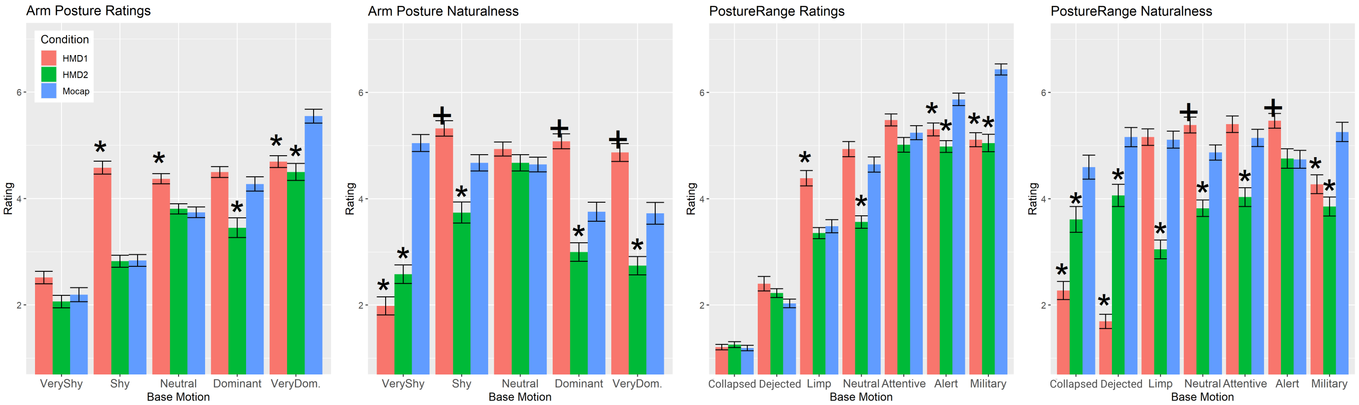


Fig. 5: Social Signal and Naturalness ratings for posture scales. * indicates HMD clips that performed significantly differently from MoCap. + indicates HMD clips that performed significantly better in naturalness than MoCap. Error bars show standard error.

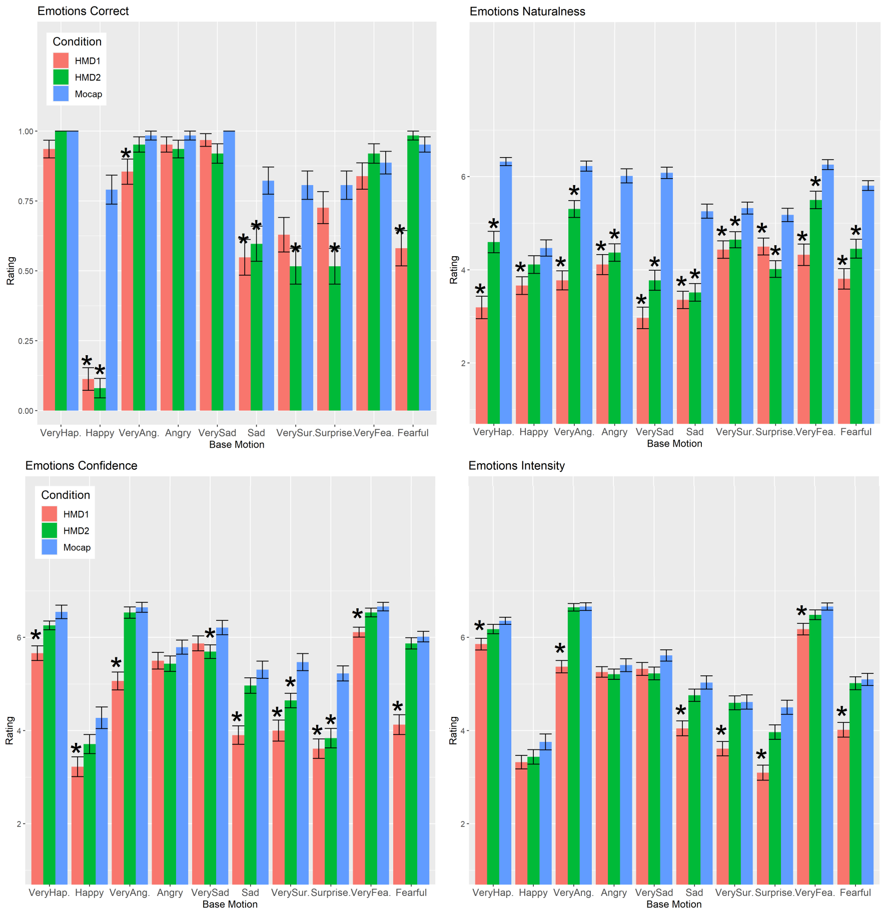


Fig. 6: Emotion correctness, Naturalness, Confidence, and Intensity rating results. * indicates HMD clips that performed significantly differently from MoCap.

Table 2: Summary of Social Signal and Naturalness ratings in comparison of HMD1 and HMD2 with MoCap. HMD1 and HMD2 diff refer to stimuli counts that are rated significantly different than MoCap. HMD1 and HMD2 worse/better, refer to stimuli counts rated significantly worse/better in the Naturalness rating than MoCap. No HMD2 stimuli were rated significantly more natural than MoCap. Cohen's d is used as a measure of effect size. d_T denotes the total and d_M the mean values.

		Social Signal Ratings						**Naturalness Ratings**								
Category	# Clips	HMD1 diff	d_T	d_M	HMD2 diff	d_T	d_M	HMD1 worse	d_T	d_M	HMD1 better	d_T	d_M	HMD2 worse	d_T	d_M
Attracted vs. Repulsed	5	1	.68	.68	1	.68	.68	5	7.43	1.49	0	0		5	5.89	1.18
Dominant vs. Shy	5	5	3.31	.66	2	1.17	.58	4	4.82	1.21	0	0		5	6.36	1.27
Energetic vs. Tired	5	1	.93	.93	1	.58	.58	5	7.38	1.48	0	0		5	7.98	1.60
Interested vs. Disinterested	5	4	2.84	.71	3	1.93	.64	5	4.80	.96	0	0		5	7.09	1.42
Positive vs. Negative	5	3	4.16	1.39	1	.51	.51	4	5.64	1.41	0	0		5	4.48	.90
Relaxed vs. Nervous	5	4	3.41	.85	4	5.12	1.28	4	3.93	.98	0	0		5	5.60	1.12
Arm Posture	5	3	3.62	1.21	2	1.56	.78	1	2.35	2.35	3	2.37	.79	4	3.74	.93
Posture Range	7	3	2.87	.96	3	3.34	1.11	3	4.94	1.65	2	1.05	.52	6	5.50	.92
Happiness	2	1	1.84	1.84	1	2.03	2.03	2	2.76	1.38	0	0		1	1.26	1.26
Anger	2	1	.48	.48	0	0		2	3.20	1.60	0	0		2	2.01	1.01
Sadness	2	1	.61	.61	1	.51	.51	2	3.60	1.80	0	0		2	2.97	1.49
Surprise	2	0	0		2	1.28	.64	2	1.23	.62	0	0		2	1.48	.74
Fear	2	1	.97	.97	0	0		2	2.85	1.43	0	0		2	1.70	.85
Totals	52	**28**	**25.72**	.93	**21**	**18.70**	.79	**41**	**54.93**	1.43	**5**	**3.41**	.66	**49**	**56.06**	1.15
Percent with Error		**53.85**			**40.38**			**78.85**			**9.62**			**94.23**		

this does not account for instances where HMD1 was rated better than motion capture. HMD1 was worse than motion capture fewer times (i.e., 41 vs. 49). Directly comparing the two, HMD1 was significantly better than HMD2 in 17 cases ($d_T = 15.51$), and HMD2 was better in 11 ($d_T = 9.76$).

The Emotion survey included two additional questions: one to rate participants' confidence in their answers and the second to rate the intensity of the motion. Both of these were lower for the HMD reconstructions. Fitting a cumulative link model, the intended emotion ($\chi^2[9] = 617.6$, $p < .0001$), reconstruction algorithm ($\chi^2[2] = 225.7$, $p < .0001$), and interaction of these factors ($\chi^2[18] = 89.7$, $p < .0001$)

were all significant. Participants were less confident in HMD1 than motion capture for: *Very Happy*, *Happy*, *Very Angry*, *Sad*, *Very Surprised*, *Surprised*, *Very Fearful*, and *Fearful*. Participants were less confident in HMD2 than motion capture for: *Very Happy*, *Very Sad*, *Surprised*, and *Very Surprised*. In addition, participants were less confident in HMD1 than in HMD2 for: *Very Happy*, *Very Angry*, *Sad*, *Very Fearful*, and *Fearful*; however, HMD2 was never significantly worse than HMD1.

Similarly, for Intensity rating, fitting a cumulative link model shows a significant main effect for the intended emotion ($\chi^2[9] = 933.7$, $p < .0001$), reconstruction algorithm ($\chi^2[2] = 162.9$, $p < .0001$), and their interaction ($\chi^2[18] = 84.5$, $p < .0001$). HMD1 was perceived as less intense than motion capture for: *Very Happy*, *Very Angry*, *Sad*, *Surprised*, *Very Surprised*, *Very Fearful*, and *Fearful*. HMD2 was never perceived as significantly less intense than motion capture. HMD1 was perceived as significantly less intense than HMD2 for *Very Angry*, *Sad*, *Very Surprised*, *Surprised*, and *Fearful*.

Overall, there is evidence that the HMD conditions often led to lower confidence and less perceived emotional intensity. In both cases, HMD1 performed worse than HMD2.

4.3 Qualitative Results

We conducted qualitative analysis by reviewing the corresponding stimuli when there were significant differences in the ratings. Because we matched stimuli across the three motion conditions, with camera footage of each motion capture session, we can examine the clips and make reasonable postulates about which feature change in the motion is leading to a change in the particular Social Signal or Naturalness ratings. Two of our researchers listed out their observations and discussed till reaching a mutual agreement on plausible reasons behind each significantly different rating. Qualitative observations are discussed below.

Social signals in HMD1 and HMD2 were degraded due to three classes of errors: arm tracking, posture reconstruction, and floating. Please see the accompanying video for examples. Several forms of arm errors were observed. Gestures at the character's side or below the waist were frequently not captured. The arm swivel angle, which controls the distance between the elbows and the character's side, was not accurate, resulting in incorrect pose readings when showing dominance. Shoulder errors also occurred. If the hands were too close to the body, they were not tracked well. This impacted a self-hug in a sad motion. There were pops in the arm motions for HMD1 and sometimes jerkiness in the arm motion for HMD2. This made the character read as less relaxed. Finally, the actor expressed surprise with subtle torso arousal and one arm raised in front of the chest. HMD1 missed the arm posture but showed sudden torso movements, while HMD2 captured the arm with slight positional errors, making the expression seem more like anger than surprise, resulting in a better Social Signal rating in HMD1 than HMD2.

Posture errors occurred in the sagittal and coronal planes, as well as in the lower body. Failure to accurately reconstruct sagittal posture adjustments resulted in the *Limp* posture and the *Very Tired* clip not being read clearly. High-frequency coronal sway in the shoulder and spine was the basis of the *Happy* motion, and neither HMD algorithm could track it, causing the *Happy* stimulus on the emotion scale to be scored much lower in social signal correctness for both HMD1 and HMD2 compared to MoCap. A hip tilt and turn out of a leg was often used to adopt a more casual pose, and HMD tracking could not reconstruct this.

A final source of error stems from floating and overall body stance, an area where HMD1 performs more poorly than HMD2, although it is an issue for both. In the Interested clip, HMD1 reads a forward lean as a forward translation, reducing the sense of interest. It treats a headshake in the Disinterested clip as a rotation of the body, resulting in the negation of the headshake no longer reading, and the clip being rated as more interesting than the original motion capture.

5 Discussion

This section discusses how Social Signal and Naturalness ratings jointly characterize differences between HMD1, HMD2, and MoCap. We show that HMD2 preserved social meaning more accurately while both HMD methods were frequently rated less natural than MoCap, and we relate these outcomes to recurrent failure modes (e.g., arm swivel/shoulder errors, posture reconstruction issues, and floating/stance artifacts) with implications for evaluation and tracking design.

- **RQ1:** Results showed that HMD2 conveyed social information better than HMD1, relative to MoCap as ground truth.

HMD2 scores better (i.e., lower) on both the significantly different stimuli count and total effect size. Across all developed scales, HMD2 had a lower or equal count of significant differences compared to HMD1, except *Surprise* in the emotion scale. Cases where HMD2 performed better than HMD1 included *Dominant* vs. *Shy*, *Interested* vs. *Disinterested*, *Positive* vs. *Negative*, *Arm Posture*, *Anger*, and *Fear*. HMD2 also performed better on the Intensity and Confidence ratings in the emotion scale than HMD1. This is consistent with our expectations, given that HMD2 with IOBT utilizes more input from the headset cameras that can capture more nuances in the body movement. This result aligns with the work of Castillo and Neff [7], who emphasize that subtle differences in hand shape and motion intensity shift the perceived valence of a motion. Relying on 3pt tracking, HMD1 frequently exhibited lower intensity and less accurate arm and shoulder positions compared to HMD2, which led to a more significant shift in delivering the correct encoded social signals.

- **RQ2:** Results showed that both HMD1 and HMD2 were rated less natural than MoCap, with HMD1 showing higher ratings in naturalness than HMD2.

The Naturalness ratings did not show stronger performance for HMD2. HMD1 outperforms on the count metric (i.e., 41 to 49), and the total effect sizes are nearly identical (see Table 2). HMD1 was even rated as more natural than MoCap for three motions in the *Arm Posture* and two in the *Posture Range* scale. In direct comparisons, HMD1 was often rated as more natural than HMD2 (i.e., 17 to 11). Specifically, in *Dominant* vs. *Shy*, *Interested* vs. *Disinterested*, *Relaxed* vs. *Nervous*, *Arm Posture*, and *Posture Range* scales. HMD2 only outperformed HMD1 in the *Positive* vs. *Negative* scale, see Table 2 in the supplemental documentation. While Naturalness ratings did reflect the overall drop in quality of HMD-based tracking compared to MoCap, participants' ratings between HMD1 and HMD2 showed informative distinctions that highlighted specific differences in tracking fidelity of the two motion reconstruction algorithms. HMD1 performed well on naturalness at times because it missed parts of the real motion that were out of view and ended up generating very little motion and a relatively natural resting pose. This was observed in all the cases where it outscored motion capture. The pattern of jerkiness also varied, with less frequent but larger jerks occurring with HMD1 and more constant jerkiness occurring in some motions with HMD2. This finding echoed Ferstl et al.'s [13] where they evaluated perceived naturalness on HMD-based motion errors resulting from hand-tracking loss and found that sudden pops and lack of smoothness were noticeable to observers, lowering the Naturalness rating. Moreover, some participants commented that they based their ratings on whether the movements were smooth and whether they felt that the motions were realistic reactions that a person would have. The latter aligns with Ren et al.'s work on investigating synthetic motion naturalness, where they stated that *"motions that we have seen repeatedly are judged natural, whereas motions that happen very rarely are not"* [43]. This suggests that Naturalness ratings might represent a conflation of (1) whether the motion is showing less jerkiness and resembling human natural movement, and (2) whether the motion is something commonly done in social settings.

- **RQ3:** Results implied that HMD2 has better Social Signal ratings than HMD1, while HMD1 has better Naturalness ratings than HMD2.

The divergence of Social Signal ratings and Naturalness ratings we observed in the study suggests that perceived social signals and naturalness are two separable constructs, and both can provide valuable insight for motion evaluation. Different motion reconstruction solutions might encounter varying strengths and weaknesses depending on the input tracking data, algorithms, and social cues presented. Our results showed that although HMD2 (i.e., IOBT) performed better in conveying social signals, it was rated less natural compared to HMD1 due to the introduced jerkiness from additional body-tracking inputs. In contrast, HMD1 lost certain gestures and body movements, resulting in smoother and more natural motion, but was not able to deliver social signals accurately. In summary, motions that are less natural might still provide higher correctness in social cues. Meanwhile, motions that are more natural do not guarantee accurate social signals. Understanding the reasons behind such ratings requires qualitative investigations on the generated clips. Similar to the proposed evaluation metrics in the GENEA challenge by Kucherenko et al. [28, 29], which was adopted in further evaluation studies [1, 55], we should separate evaluation on human-likeness (i.e., naturalness) and communication appropriateness to better understand the pros and cons of each tracking reconstruction method. This is especially important as we move to GenAI systems for synthetic motions. These systems will be increasingly capable of creating natural motion, but if we do not also measure the content being delivered, they may not lead to the authentic communication desired.

- **RQ4:** With both quantitative and qualitative analysis, we were able to identify why HMD1 and HMD2 produce weaker ratings on particular social signal scales.

Our study design facilitated qualitative analysis of the motion stimuli, looking into why HMD1 or HMD2 failed on certain scales. If a stimulus's Social Signal rating fails on a particular scale, such as *Happy* or *Surprise*, we can easily perform a qualitative evaluation of the stimulus and postulate what might be the issue causing the inaccurate rating. Results showed that three main errors: arm tracking, posture reconstruction, and floating were the most prominent causes in the degradation of Social Signal ratings. The same applies when a stimulus fails on the Naturalness ratings. For instance, we concluded that HMD1 performed more naturally than HMD2, and even more than MoCap, in the posture scale simply because it damped certain motions and made the whole stimulus smoother. However, since Naturalness ratings might be affected by both motion quality and commonality, as mentioned in RQ2, it could be more challenging to justify in certain clips. Overall, the design of our evaluation scales with collected ratings and qualitative observations helps inform the design of the motion reconstruction algorithms, identify their deficiencies, and better adapt to motions seen in social settings.

6 Limitations

While our study has certain limitations that should be acknowledged, these do not undermine the validity of our results. Instead, they point to areas for improvement in future research. First, although our designed scales aimed to cover the majority of social motions, emblems and gestures were not part of our design focus. Emblem reconstruction strongly relies on HMD hand tracking quality, and our study aimed to focus more on HMD body tracking. However, the accuracy of social signals with emblems is also essential in evaluating HMD-based motion reconstruction. Second, our design of one motion at each scale point precluded our ability to generalize if the reconstruction algorithm always fails or succeeds for a particular social signal. An alternative approach is to test all motions that passed the selection study in one VR session, but this is time-consuming and can easily introduce fatigue in our study participants. Furthermore, although we attempted to control for actor-related encoding effects by recruiting a diverse pool of actors, the selection study prioritized stimuli that best spanned across the rating ranges, which may result in a final benchmark that is not gender-balanced and may therefore introduce gender effects that were not systematically examined. Lastly, we only tested our designed scales with two forms of HMD body tracking in our study. The existing challenges of lower-body tracking with HMD devices prohibit reliable knee, hip, and feet poses with HMD1 and HMD2, limiting their performance in full-body reconstructions. Additionally, auditory data is a critical factor in signal interpretation. Our work focuses on testing the designed benchmark without interaction of audio; however, evaluation of HMD-based tracking reconstructions with audio might provide further insight into the results in future studies.

7 Conclusions and Future Work

Tracking from headsets is a challenging task. While our findings showed that both tracking reconstructions are impressive in many ways, they still fall short of motion capture in both our Social Signal and Naturalness ratings. The results for the social signal measures suggest the possibility of miscommunication or diminished communication. However, besides our interesting findings, much future work remains. In real applications, body motion is only one of a set of signals that includes dialog, tone of voice, and the appearance of the avatar. It would be worthwhile to investigate the impact of motion relative to these other social signals. The dataset could also support interesting future investigations into the perception of social signals. Given the data for two clips (e.g., HMD1 and MoCap) that are rated differently on a particular social signal, it would be possible to create in-between clips that adjust specific features in the motion. These could be used in new experiments to test, for example, how much motion jitter versus posture changes contributed to a particular shift in social signals. All in all, we hope that this work will support future evaluation efforts. It sets a benchmark for HMD tracking reconstruction, allowing future algorithms to be compared against using these motion data and results. More generally, it suggests a new approach to animation evaluation that prioritizes measuring the messages conveyed through motion.

Acknowledgments

We thank Vivian Lo, Steffi Liem, Callie Waldman, Nicky He, Michelle Hill, Tiffany Popov, Jenna Zarate, and Nelson Tam for their assistance, as well as Doug Roble and Yuting Ye for valuable feedback. We are also very grateful to all of the talented actors for providing such rich performances.

References

[1] L. Abel, V. Colotte, and S. Ouni. Towards interpretable co-speech gestures synthesis using stargate. In *Companion Proceedings of the 26th International Conference on Multimodal Interaction*, pp. 138–146, 2024. 9

[2] A. Adkins, A. Normoyle, L. Lin, Y. Sun, Y. Ye, M. Di Luca, and S. Jörg. How important are detailed hand motions for communication for a virtual character through the lens of charades? *ACM Transactions on Graphics*, 42(3):1–16, 2023. 3

[3] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1(1):71–81, 2009. 2

[4] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of statistical software*, 67:1–48, 2015. 5

[5] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995. 5

[6] R. Canales, A. Normoyle, Y. Sun, Y. Ye, M. D. Luca, and S. Jörg. Virtual grasping feedback and virtual hand ownership. In *ACM Symposium on Applied Perception 2019*, pp. 1–9, 2019. 3

[7] G. Castillo and M. Neff. What do we express without knowing? emotion in gesture. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 702–710, 2019. 3, 8

[8] S. Castillo, K. Legde, and D. W. Cunningham. The semantic space for motion-captured facial expressions. *Computer Animation and Virtual Worlds*, 29(3-4):e1823, 2018. 3

[9] R. H. B. Christensen. Analysis of ordinal data with cumulative link models—estimation with the r-package ordinal. *R-package version*, 28, 2015. 5

[10] R. H. B. Christensen. Cumulative link models for ordinal regression with the r package ordinal. *Submitted in J. Stat. Software*, 2018. 5

[11] F. Danieau, T. Lopez, N. Mollet, B. Leroy, O. Dumas, and J.-F. Vial. Enabling embodiment and interaction in omnidirectional videos. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 697–702. IEEE, 2017. 3

[12] J. C. Eubanks, A. G. Moore, P. A. Fishwick, and R. P. McMahan. The effects of body tracking fidelity on embodiment of an inverse-kinematic avatar for male participants. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 54–63. IEEE, 2020. 3

[13] Y. Ferstl, R. McDonnell, and M. Neff. Evaluating study design and strategies for mitigating the impact of hand tracking loss. In *ACM Symposium on Applied Perception 2021*, pp. 1–12, 2021. 2, 3, 8

[14] Y. Ferstl, S. Thomas, C. Guiard, C. Ennis, and R. McDonnell. Human or robot? investigating voice, appearance and gesture motion realism of conversational social agents. In *Proceedings of the 21st ACM international conference on intelligent virtual agents*, pp. 76–83, 2021. 2

[15] S. Han, B. Liu, R. Wang, Y. Ye, C. D. Twigg, and K. Kin. Online optical marker-based hand tracking with deep labels. *Acm transactions on graphics (tog)*, 37(4):1–10, 2018. 4

[16] J. Harrison, R. A. Rensink, and M. Van De Panne. Obscuring length changes during animated motion. *ACM Transactions on Graphics (TOG)*, 23(3):569–573, 2004. 2

[17] F. Herrera, S. Y. Oh, and J. N. Bailenson. Effect of behavioral realism on social interactions inside collaborative virtual environments. *Presence*, 27(2):163–182, 2020. 2

[18] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2

[19] J. Hodgins, S. Jörg, C. O'Sullivan, S. I. Park, and M. Mahler. The saliency of anomalies in animated human characters. *ACM Transactions on Applied Perception (TAP)*, 7(4):1–14, 2010. 2

[20] J. K. Hodgins, J. F. O'Brien, and J. Tumblin. Perception of human motion with different geometric models. *IEEE Transactions on Visualization and Computer Graphics*, 4(4):307–316, 1998. 2

[21] L. Hoyet, K. Ryall, K. Zibrek, H. Park, J. Lee, J. Hodgins, and C. O'sullivan. Evaluating the distinctiveness and attractiveness of human motions on realistic virtual bodies. *ACM Transactions on Graphics (TOG)*, 32(6):1–11, 2013. 2

[22] E. Jain, L. Anthony, A. Aloba, A. Castonguay, I. Cuba, A. Shaw, and J. Woodward. Is the motion of a child perceivably different from the motion of an adult? *ACM Transactions on Applied Perception (TAP)*, 13(4):1–17, 2016. 2

[23] P. Jonell, Y. Yoon, P. Wolfert, T. Kucherenko, and G. E. Henter. Hemvip: Human evaluation of multiple videos in parallel. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pp. 707–711, 2021. 2

[24] S. Kenny, N. Mahmood, C. Honda, M. J. Black, and N. F. Troje. Effects of animation retargeting on perceived action outcomes. In *Proceedings of the ACM Symposium on Applied Perception*, pp. 1–7, 2017. 2

[25] S. Kenny, N. Mahmood, C. Honda, M. J. Black, and N. F. Troje. Perceptual effects of inconsistency in human animations. *ACM Transactions on Applied Perception (TAP)*, 16(1):1–18, 2019. 2

[26] A. Koilias, C. Mousas, and C.-N. Anagnostopoulos. The effects of motion artifacts on self-avatar agency. In *Informatics*, vol. 6, p. 18. MDPI, 2019. 2

[27] E. Kokkinara and R. McDonnell. Animation realism affects perceived character appeal of a self-virtual face. In *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games*, pp. 221–226, 2015. 2

[28] T. Kucherenko, R. Nagy, Y. Yoon, J. Woo, T. Nikolov, M. Tsakov, and G. E. Henter. The genea challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pp. 792–801, 2023. 2, 9

[29] T. Kucherenko*, P. Wolfert*, Y. Yoon*, C. Viegas, T. Nikolov, M. Tsakov, and G. E. Henter. Evaluating gesture generation in a large-scale open challenge: The genea challenge 2022. *ACM Transactions on Graphics*, 43(3):1–28, 2024. 2, 9

[30] K. Liu, J. Tolins, J. E. F. Tree, M. Neff, and M. A. Walker. Two techniques for assessing virtual agent personality. *IEEE Transactions on Affective Computing*, 7(1):94–105, 2015. 2

[31] I. Meta Platforms. Movement sdk for unity-overview, Dec 2024. 4

[32] I. Meta Platforms. Momentum: A library for human kinematic motion and numerical optimization solvers to apply human motion, 2025. 4

[33] I. Meta Platforms. Move body tracking with unity, May 2025. 4

[34] I. Meta Platforms. Vrs overview, 2025. 4

[35] M. Neff. Lessons from the arts: what the performing arts literature can teach us about creating expressive character movement. *Nonverbal Communication in Virtual Worlds: Understanding and Designing Expressive Characters*, pp. 123–148, 2014. 3

[36] M. Neff, N. Toothman, R. Bowmani, J. E. F. Tree, and M. A. Walker. Don't scratch! self-adaptors reflect emotional stability. In *International Workshop on Intelligent Virtual Agents*, pp. 398–411. Springer, 2011. 3

[37] M. Neff, Y. Wang, R. Abbott, and M. Walker. Evaluating the effect of gesture and language on personality perception in conversational agents. In *International Conference on Intelligent Virtual Agents*, pp. 222–235. Springer, 2010. 3

[38] S. Nyatsanga, D. Roble, and M. Neff. The impact of avatar retargeting on pointing and conversational communication. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 2, 3

[39] J. L. Ponton, H. Yun, A. Aristidou, C. Andujar, and N. Pelechano. Sparseposer: Real-time full-body motion reconstruction from sparse data. *ACM Transactions on Graphics*, 43(1):1–14, 2023. 1

[40] P. S. Reitsma and C. O'Sullivan. Effect of scenario on perceptual sensitivity to errors in animation. *ACM Transactions on Applied Perception (TAP)*, 6(3):1–16, 2009. 2

[41] P. S. Reitsma and N. S. Pollard. Perceptual metrics for character animation: sensitivity to errors in ballistic motion. In *ACM SIGGRAPH 2003 Papers*, pp. 537–542. 2003. 2

[42] R. Rekik, S. Wuhrer, L. Hoyet, K. Zibrek, and A.-H. Olivier. A survey on realistic virtual human animations: Definitions, features and evaluations. In *Computer Graphics Forum*, vol. 43, p. e15064. Wiley Online Library, 2024. 2, 3

[43] L. Ren, A. Patrick, A. A. Efros, J. K. Hodgins, and J. M. Rehg. A data-driven approach to quantifying natural human motion. *ACM Transactions on Graphics (TOG)*, 24(3):1090–1097, 2005. 8

[44] T. Shawn. *Every Little Movement: A Book about Francois Delsarte*. Dance Horizons, Inc., New York, second revised ed., 1963. 3

[45] M. Slater, A. Sadagic, M. Usoh, and R. Schroeder. Small-group behavior in a virtual and real environment: A comparative study. *Presence: Teleoperators & Virtual Environments*, 9(1):37–51, 2000. 1

[46] H. J. Smith and M. Neff. Understanding the impact of animated gesture performance on personality perceptions. *ACM Transactions on Graphics (TOG)*, 36(4):49, 2017. 3

[47] H. J. Smith and M. Neff. Communication behavior in embodied virtual reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 289. ACM, 2018. 1

[48] J.-P. Tauscher, M. Mustafa, and M. Magnor. Comparative analysis of three different modalities for perception of artifacts in videos. *ACM Transactions on Applied Perception (TAP)*, 14(4):1–12, 2017. 2

[49] A. Thaler, S. Pujades, J. K. Stefanucci, S. H. Creem-Regehr, J. Tesch, M. J. Black, and B. J. Mohler. The influence of visual perspective on body size estimation in immersive virtual reality. In *ACM Symposium on Applied Perception 2019*, pp. 1–12, 2019. 2

[50] N. Toothman and M. Neff. The impact of avatar tracking errors on user experience in vr. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 756–766. IEEE, 2019. 2, 3

[51] J. Tseng, R. Castellon, and K. Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 448–458, 2023. 2

[52] J. Voas, Y. Wang, Q. Huang, and R. Mooney. What is the best automated metric for text to motion generation? In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023. 2

[53] C. Wallraven, M. Breidt, D. W. Cunningham, and H. H. Bülthoff. Evaluating the perceptual realism of animated facial expressions. *ACM Transactions on Applied Perception (TAP)*, 4(4):1–20, 2008. 3

[54] T. Waltemate, I. Senna, F. Hülsmann, M. Rohde, S. Kopp, M. Ernst, and M. Botsch. The impact of latency on perceptual judgments and motor performance in closed-loop interaction in virtual reality. In *Proceedings of the 22nd ACM conference on virtual reality software and technology*, pp. 27–35, 2016. 3

[55] A. W. Werner, J. Beskow, and A. Deichler. Gesture evaluation in virtual reality. In *Companion Proceedings of the 26th International Conference on Multimodal Interaction*, pp. 156–164, 2024. 9

[56] P. Wisessing, K. Zibrek, D. W. Cunningham, J. Dingliana, and R. McDonnell. Enlighten me: Importance of brightness and shadow for character emotion and appeal. *ACM Transactions on Graphics (TOG)*, 39(3):1–12, 2020. 2

[57] L. Wöhler, S. Castillo, and M. Magnor. Personality analysis of face swaps: can they be used as avatars? In *Proceedings of the 22nd ACM international conference on intelligent virtual agents*, pp. 1–8, 2022. 3

[58] P. Wolfert, J. M. Girard, T. Kucherenko, and T. Belpaeme. To rate or not to rate: Investigating evaluation methods for generated co-speech gestures. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pp. 494–502, 2021. 2

[59] P. Wolfert, N. Robinson, and T. Belpaeme. A review of evaluation practices of gesture generation in embodied conversational agents. *IEEE Transactions on Human-Machine Systems*, 52(3):379–389, 2022. 2

[60] G. Yamac, C. O'Sullivan, and M. Neff. Understanding the impact of visual and kinematic information on the perception of physicality errors. *ACM Transactions on Applied Perception*, 2024. 2

[61] Y. Ye, L. Liu, L. Hu, and S. Xia. Neural3points: Learning to generate physically realistic full-body motion for virtual reality users. In *Computer Graphics Forum*, vol. 41, pp. 183–194. Wiley Online Library, 2022. 1

[62] Y. Yoon, B. Cha, J.-H. Lee, M. Jang, J. Lee, J. Kim, and G. Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020. 2

[63] H. Yun, J. L. Ponton, C. Andujar, and N. Pelechano. Animation fidelity in self-avatars: Impact on user performance and sense of agency. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 286–296. IEEE, 2023. 3

[64] K. Zibrek, E. Kokkinara, and R. McDonnell. The effect of realistic appearance of virtual characters in immersive environments-does the character's personality play a role? *IEEE transactions on visualization and computer graphics*, 24(4):1681–1690, 2018. 3