Reality Check: How Avatar and Face Representation Affect the Perceptual Evaluation of Synthesized Gestures

HAOYANG DU, Technological University Dublin, Ireland
YINGHAN XU, Trinity College Dublin, Ireland
JOHN DINGLIANA, Trinity College Dublin, Ireland
BRIAN KEEGAN, Technological University Dublin, Ireland
RACHEL MCDONNELL, Trinity College Dublin, Ireland
CATHY ENNIS, Maynooth University, Ireland



Fig. 1. The seven avatar representations evaluated in this study. From left to right: Gaussian Avatar (Reconstruction-based Avatars), Deploy-Hi and Deploy-Lo (high-fidelity and lightweight deployable assets), TexSMPL-X and UntexSMPL-X and Mann (standard research baselines), and Stick (minimalist kinematic visualization baseline). We explore how avatar and face visualization shapes perceptual evaluations of co-speech gestures in a series of 3 experiments.

The capacity to create realistic virtual humans has progressed significantly, and such characters can be found in many applications across entertainment, education and health. As an essential element of interactive virtual humans, speech-driven 3D gesture generation still depends heavily on perceptual evaluation, yet studies often vary avatar appearance and facial presentation when judging the generated motions. Prior work suggests these visual choices can bias motion judgments, but controlled evidence remains limited. We address this gap with controlled evaluations of co-speech gestures across motion sources, spanning seven representative avatar renderings used in contemporary research and application pipelines. Our results show that avatar and face presentation systematically shift perceptual judgments, and we provide recommendations for benchmarking gesture synthesis as well as for deploying virtual humans in human-facing applications.

CCS Concepts: • **Computing methodologies** → **Animation**; • **Human-centered computing** → **Empirical studies in HCI**.

ACM Reference Format:
Haoyang Du, Yinghan Xu, John Dingliana, Brian Keegan, Rachel McDonnell, and Cathy Ennis. 2026. Reality Check: How Avatar and Face Representation Affect the Perceptual Evaluation of Synthesized Gestures. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '26), July 19–23, 2026, Los Angeles, CA, USA*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3799902.3811161

Authors' Contact Information: Haoyang Du, Technological University Dublin, Dublin, Ireland, D23128268@mytudublin.ie; Yinghan Xu, Trinity College Dublin, Dublin, Ireland, yixu@tcd.ie; John Dingliana, Trinity College Dublin, Dublin, Ireland, john.dingliana@tcd.ie; Brian Keegan, Technological University Dublin, Dublin, Ireland, brian.x.keegan@TUDublin.ie; Rachel McDonnell, Trinity College Dublin, Dublin, Ireland, ramcdonn@tcd.ie; Cathy Ennis, Maynooth University, Maynooth, Ireland, Cathy.Ennis@mu.ie.



This work is licensed under a Creative Commons Attribution 4.0 International License.
SIGGRAPH Conference Papers '26, Los Angeles, CA, USA
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2554-8/26/07
https://doi.org/10.1145/3799902.3811161

1 Introduction

As conversational agents expand into social and professional roles [Cui and Liu 2023], the automatic generation of co-speech gestures has emerged as a key aspect for natural interaction [Ali et al. 2025]. This field is undergoing a significant data-driven transformation; large-scale multi-modal datasets now allow models to approach the realism of motion capture [Agrawal et al. 2025]. As a result, current evaluation frameworks must evolve to detect the subtle remaining gaps between generated motion and human-captured motion.

Avatar visualization is a key factor in this evaluation, because an avatar's visual realism shapes expectations and tolerance for motion artifacts [Rekik et al. 2024]. However, current studies use widely different avatar types, which can bias outcomes and limit cross-paper comparability [Nagy et al. 2025]. At the same time, avatar modeling itself is moving toward reconstruction-based rendering (e.g., Gaussian-based representations), making it increasingly important to understand how avatar choice shapes motion judgments and to support more reliable evaluation across studies.

To address this problem, we conduct three experiments to examine how avatar appearance shapes motion judgments. Because facial cues can distract attention from gesture [Hessels 2020], we first identify a face visualization that minimizes facial confounds during gesture evaluation. Using this visualization, we then test

seven avatar styles commonly used in application pipelines and research to assess how avatar appearance modulates perceived motion quality. Finally, we validate these findings using muted speech to isolate the influence of visual motion.

Drawing on the results of our three experiments, we offer the following practical recommendations. For benchmarking, we recommend separating facial and body cues and blurring the face during evaluation to reduce facial confounds, avoid perceptual mismatches, and minimize distraction. We further suggest reconstruction-based avatars (e.g., Gaussian) for evaluation, as it consistently shows the clearest separation between generated and human-captured motion across our evaluation measures. For deployment, however, we caution against using Gaussian Avatar when motion is unreliable: pairing visually complex appearance with natural or incongruent motion can reduce comprehension and increase negative affect, making simpler avatar representations a safer choice.

2 Related Work

Gestures produced alongside speech help convey meaning and structure discourse in human conversation [Clough and Duff 2020], motivating automatic gesture generation for virtual agents. The field has progressed from rule-based mappings and motion-clip retrieval to data-driven models trained on paired speech–motion datasets, and with the availability of large multimodal corpora, deep learning has become dominant; modern generative approaches have improved realism, diversity, and semantic coherence [Nyatsanga et al. 2023]. Evaluation of synthetic gesture generation typically combines objective metrics and subjective human judgments, yet there is still no field-wide standard [Nyatsanga et al. 2023]: objective measures can capture similarity to reference motion but often fail to reflect perceptual qualities such as naturalness and speech–gesture appropriateness [Crnek et al. 2025; Haque et al. 2025], making subjective evaluation indispensable; however, studies vary substantially in rating instruments, stimuli, baselines, and display conditions, limiting cross-paper comparability [Wolfert et al. 2022].

To address this, the GENEA Challenges introduced a community benchmark with shared evaluation protocols [Nagy et al. 2025], which separates subjective assessment into motion human-likeness and speech–gesture alignment, often using mismatching paradigms to reduce confounding effects [Kucherenko et al. 2023, 2024]. Under this framework, many gesture-generation systems produce highly natural motion yet exhibit weak appropriateness to speech, with alignment accuracy often dropping to near-chance under mismatched conditions [Nagy et al. 2025]. More recently, large-scale dyadic modeling has demonstrated that mocap-level alignment is achievable [Agrawal et al. 2025], suggesting that we may need more sensitive evaluation tools along both motion human-likeness and alignment to detect fine-grained improvements.

Beyond protocols and rating instruments, avatar visualization can bias perceived motion quality [Zibrek and McDonnell 2014]. In particular, the face can exert strong visual dominance, drawing attention away from gestures during evaluation [Hessels 2020]. To mitigate this effect, GENEA uses a masked face, while other studies have used blurring [McDonnell et al. 2008] or minimally animated faces [Du et al. 2025]. However, these manipulations can also influence perception: static faces are perceived as less natural [Kullmann et al. 2023], and masked faces can cause discomfort [Wöhler et al. 2024]. In contrast, preserving active facial cues increases perceived realism and appeal, while also contributing to improved audiovisual integration [Amadou et al. 2023; Holler et al. 2014].

Inconsistent avatar visualization may also hinder subtle comparison across studies. Prior work spans stick figures [Cheng et al. 2025; Qi et al. 2024], cartoon characters [Chen et al. 2025; Zhang et al. 2024], untextured meshes [Chhatre et al. 2024; Liu et al. 2024], and high-fidelity human avatars [Ng et al. 2024], limiting meaningful cross-study comparison [Wolfert et al. 2022]. Although recent work has moved toward more controlled setups (e.g., SMPL-X mesh characters with masked face) [Nagy et al. 2025], there is still no systematic investigation of how avatar visualization affects perceived gesture quality. As a result, the role of avatar visualization in gesture evaluation remains largely unknown.

More broadly, avatar appearance shapes how virtual motion is perceived, and visual realism can bias judgments beyond motion itself. Increasing anthropomorphism can make identical motions seem less biological [Chaminade et al. 2007], while photo-textured humans elicit lower tolerance for motion anomalies than stylized characters, implying higher fidelity raises expectations [McDonnell et al. 2012]. Higher animation quality and appearance realism further increase perceived realism and comprehension [Adkins et al. 2023; Amadou et al. 2023]. Recent work shows that high-fidelity avatars increase sensitivity in gesture evaluation compared to mesh-based renderings, but does not disentangle facial cues or avatar representation from motion quality itself [Ng et al. 2024].

The interaction between realism and movement connects to the Uncanny Valley: near-realistic characters can seem eerier when their motion is unnatural [Rekik et al. 2024]. This aligns with the Mismatch Hypothesis, where inconsistencies across channels trigger prediction errors and negative reactions [MacDorman and Chattopadhyay 2016]. For example, pairing a photorealistic virtual human with a synthetic voice can reduce affinity [Higgins et al. 2022]. Yet mismatch is not always harmful: people may prefer highly realistic voice and motion despite appearance conflicts, depending on cue reliability and visual expectations [Ferstl et al. 2021]. Immersion can amplify these dynamics, VR can widen perceived gaps between natural and synthetic gesture–voice pairings, making synthetic motion cues more salient [Du et al. 2025].

In summary, face and avatar visualization can bias motion judgments and may interact with cross-modal mismatch, yet controlled cross-avatar evidence is lacking. We therefore test whether AI-generated gesture ratings are consistent across visualization styles and whether visual presentation modulates mismatch discomfort (e.g., uncanny-valley effects) under a unified evaluation protocol.

3 Study Overview

This study comprises three experiments summarized in Table 1. First, the Face Representation experiment identifies a facial presentation method that minimizes confounds, allowing participants to focus solely on evaluating gesture quality. Next, the Avatar Representation experiment examines how different avatar representations

Table 1. Summary of experimental designs and dependent measures across the three experiments.

Exp.	Design	Between-subjects conditions	Within-subjects conditions	Measurements
Face Representation	4 × 2 mixed-factorial	Face representation: Static face, Dynamic face, Blurred face, Masked face	Motion: Mocap, Synthetic	Speech–gesture match; Anthropomorphism; Likeability; Distraction
Avatar Representation	7 × 3 mixed-factorial	Avatar representation: Gaussian, Deploy-Hi, Deploy-Lo, TexSMPL-X, UntexSMPL-X, Mann, Stick	Motion: Matched Mocap, Mismatched Mocap, Synthetic motion	Motion human-likeness; Speech–gesture match; Comprehension; Appeal; Eeriness
Speech-Muted Validation	Pairwise Comparison	Avatar representation: Gaussian, TexSMPL-X, Stick	Motion: Mocap vs Synthetic (paired trials)	Mocap-Advantage Score (MAS)

change people's perceptions of the character animations. Finally, our Speech-Muted Validation tests whether motion quality evaluation on different avatars remains consistent in a speech-free setting by removing audio cues, ensuring that judgments reflect visual motion quality rather than speech–gesture alignment.

3.1 Stimuli Creation

3.1.1 Motion Preparation. All motion and speech stimuli were drawn from the BEAT2 dataset [Liu et al. 2024]. We selected three neutral 20-second monologue clips from distinct actors to cover diverse semantic content and improve generalizability. Each clip was screened to exclude major motion artifacts (e.g., self-intersections, foot sliding, and missing frames). While prior large-scale gesture evaluations often use 10 second clips [Kucherenko et al. 2023] to balance informativeness and participant attention, we used 20s clips to provide richer contextual evidence per trial under our multi-factor design with only three source clips.

For each clip, we created two gesture conditions: **Human captured motion (Mocap)** and **Synthetic motion (Synthetic)**. Mocap used the BEAT2 SMPL-X motion directly without modification. Synthetic motion was generated by three state-of-the-art gesture generation models to improve the generalizability: HoloGest [Cheng and Huang 2025], Semantic Gesticulator [Zhang et al. 2024], and GestureLSM [Liu et al. 2025], each producing full-body motion from speech.

3.1.2 Avatar Selection and Sources. We selected seven avatar representations spanning four categories that are commonly used in both motion-generation research and real-world pipelines. First, the radiance field-based avatars: **Gaussian** refers to Gaussian avatars represented and rendered via an explicit radiance-field representation, which enables high-fidelity human avatar reconstruction and real-time animation using 3D Gaussian Splatting (3DGS [Kerbl et al. 2023]). Unlike traditional mesh-based avatars, it represents the avatar as a collection of 3D Gaussians optimized through differentiable rendering. In our study, we used MMLPHuman [Zhan et al. 2025] to train a subject from the HumanRF [Işık et al. 2023] dataset. Second, Deployable avatars: **Deploy-Hi**, created with Character Creator 4[1], represents a high-fidelity asset increasingly adopted for academic prototyping, while **Deploy-Lo**, generated with Ready Player Me[2], represents a lightweight avatar commonly seen in deployed interactive applications. Third, standard research baseline avatars: because many motion generation methods are built around parametric body models [Pavlakos et al. 2019], we include two SMPL-X mesh baselines, **TexSMPL-X** (textured) and **UntexSMPL-X** (untextured). We also include **Mann**, a humanoid mannequin used in prior benchmark studies [Kucherenko et al. 2023]. Finally, minimal kinematic visualization avatars: **Stick** provides a line-rendered skeleton, a widely used minimalist visualization that helps isolate motion perception from appearance cues.

3.1.3 Animation Flow. To enable a fair comparison across systems with different motion representations, we standardized all gesture outputs to the SMPL-X format. Motions already available in SMPL-X (e.g., HoloGest, GestureLSM, and Mocap) were used directly. For Semantic Gesticulator, we followed the GENEA conversion approach previously used for DiffuseStyleGesture outputs [Nagy et al. 2025; Yang et al. 2023], mapping its joint rotations to SMPL-X pose parameters, extracting the root translation, and aligning the coordinate system to SMPL-X. We then applied only a small set of fixed post-processing steps and normalized global orientation so that all sequences face a consistent direction.

Avatar animation followed two pipelines depending on skeletal compatibility. Native SMPL-X avatars (TexSMPL-X, UntexSMPL-X, Gaussian, Stick) were driven directly by SMPL-X parameters. For non-native avatars (Deploy-Hi, Deploy-Lo, Mann), we avoided retargeting artifacts by re-skinning them to the SMPL-X skeleton and applying the same SMPL-X joint rotations and root translation. This ensured identical kinematics across all avatar styles.

3.1.4 Final Build. All stimuli were rendered in Unity using the High Definition Render Pipeline (HDRP) with a single shared scene configuration across conditions (same camera, lighting, and post-processing). We used a fixed frontal camera to simulate a face-to-face conversational viewpoint, and a neutral visual setup to minimize incidental cues. For the Gaussian Avatar condition, we adapted SplatBus [Xu et al. 2026], which decouples the 3D Gaussian splatting renderer from Unity through GPU interprocess communication, enabling depth-aware compositing in Unity HDRP and ensuring consistent integration with all other avatar conditions.

4 Experiment 1: Face Representation

This experiment examined how different face representations influence participants' evaluations of gesture quality. As summarized in Table 1, we implemented four facial conditions using the same Deploy-Hi avatar to keep body appearance identical: **Static**, with a completely still face; **Dynamic**, with minimal lip-sync and blinks within conversational SEBR ranges (10.5–32.5 blinks/min) [Doughty 2001]; **Blurred**, which obscured facial details via a blurred material applied to mask geometry; and **Masked**, which removed facial cues using an opaque GENEA-style covering [Nagy et al. 2025]. For the

[1]https://www.reallusion.com/character-creator
[2]https://readyplayer.me



Fig. 2. Face manipulations: blurred face (left) and masked face (right).

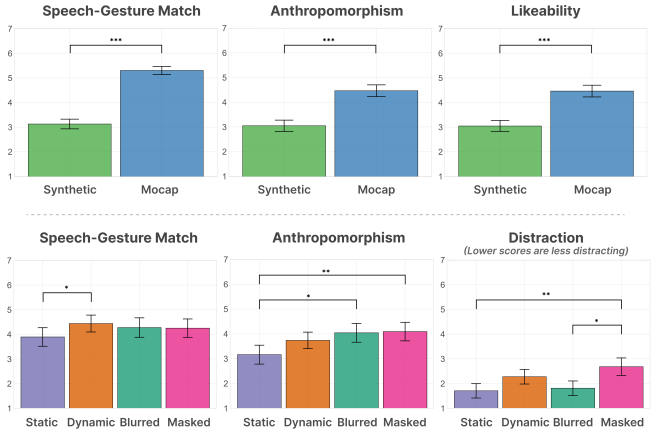


Fig. 3. Face Representation Experiment. Main effect of motion condition on various scales (top). Main effect of face condition on various scales (bottom).

blurred and masked variants, the underlying face remained completely still, while the mask geometry did not deform and followed head motion in all stimuli.

We propose the following hypotheses: **H1A**: Mocap will yield higher speech-gesture match than synthetic gestures, and the Dynamic face will further increase speech-gesture match. **H1B**: Dynamic facial animation will increase perceived anthropomorphism, while a static face will decrease it; mocap will be rated as more anthropomorphic than synthetic motion. **H1C**: Dynamic facial animation will increase perceived likeability; mocap will be rated as more likeable than synthetic motion. Unlike anthropomorphism, we do not predict that a static face will reduce likeability, as likeability does not scale linearly with anthropomorphism and can be independently driven by appearance-related factors [Bartneck et al. 2009; Zell et al. 2015]. **H1D**: Both dynamic and static facial conditions will increase perceived distraction, with no expected main effect of motion condition.

4.1 Measurements

We employed multiple validated measures to evaluate participants' responses. Participants provided ratings of gesture–speech match on a 7-point Likert scale, adapted from the GENEA evaluation protocol [Kucherenko et al. 2023], to assess whether facial alterations affect perceived speech alignment. To assess whether facial representation competed for attention during gesture evaluation, participants also rated distraction on 7-point Likert items adapted from [Wöhler et al. 2024]. To capture broader character impressions that may vary with facial visibility, we measured anthropomorphism and likeability using the corresponding Godspeed subscales [Ho and MacDorman 2010]. All items used 7-point response formats and were aggregated by averaging within each construct.

4.2 Participants & Procedure

A total of 131 participants were recruited via Prolific (63 female, 68 male; ages 21–72) and distributed evenly across four facial representation conditions ($\tilde{3}3$ per condition). Eligibility required residence in one of six English-speaking countries to reduce cross-cultural variation and a Prolific approval rate of at least 90% with at least 100 previously completed studies to ensure data quality; duplicate participation was prevented. The study was administered via Qualtrics (desktop/laptop only) at £9.50/hour. Ethics approval was obtained, informed consent collected, and participants failing a directed-response attention check were excluded.

Participants were randomly assigned to one facial representation condition and completed 6 trials (2 Motion Conditions × 3 Scenarios) presented in randomized order. For the synthetic motion condition, each participant was exposed to a single AI model across all synthetic motion trials to avoid within-participant model-specific artifacts; AI models were distributed across participants.

4.3 Results

Outcomes were analysed using mixed-design ANOVAs. Normality was assessed through Shapiro-Wilk test; outcomes violating these assumptions were analysed using the Aligned Rank Transform (ART) method [Wobbrock et al. 2011]. For brevity, Table 2 reports only significant main effects and interactions; omitted effects were non-significant. Significant effects were followed up using Bonferroni-adjusted pairwise comparisons. The main effects for Face Representation and Motion Condition are illustrated in Fig.3.

Speech-Gesture Match. A main effect of Motion Condition was found, with Mocap rated higher than Synthetic ($p < .001$). A main effect of Face Representation was also observed, with Dynamic rated higher than Static ($p < .035$).

Anthropomorphism. Mocap was rated higher than Synthetic ($p < .001$). Face Representation also had a main effect: Static was rated lower than both Blurred ($p < .018$) and Masked ($p < .010$).

Likeability. Only Motion Condition showed a main effect, with Mocap rated more likeable than Synthetic ($p < .001$).

Distraction. Face Representation affected distraction: Masked was rated more distracting than Static ($p < .003$) and Blurred ($p < .012$). Dynamic did not differ from other conditions.

4.4 Experiment 1 Discussion

Motion Condition significantly affected ratings: mocap outperformed synthetic motion on speech–gesture match, anthropomorphism, and

likeability, thereby partially supporting H1A , H1B, and H1C, and aligning with findings that natural movement improves social and communicative evaluations [Adkins et al. 2023; Du et al. 2025]. Facial representation also mattered. Consistent with H1A, speech–gesture match was higher with a Dynamic than a Static face, suggesting that minimal facial motion (e.g., lip-sync, blinking) increases perceived audiovisual integration [Krason et al. 2022; Oh Kruzic et al. 2020]. However, this may confound gesture-focused studies, as ratings may be driven by facial synchrony rather than gesture quality, making Dynamic faces a poor control for gesture evaluation.

H1B was partially supported: the Static face reduced anthropomorphism, suggesting that a frozen face paired with a moving body increased cue incongruity [Mori et al. 2012]. Contrary to H1B and H1C, the Dynamic face showed no clear advantage, likely because the reduced facial animation was insufficient to produce the expected improvement. The Masked condition was the most distracting and did not support H1D, possibly due to the opaque mask introducing a salient and unusual visual element, whereas blurring reduced facial detail without adding a conspicuous artifact. Overall, blurring offered the best low-bias baseline for isolating gesture-motion quality, so we adopted the Blurred condition in the following experiments.

5 Experiment 2: Avatar Representation

Experiment 2 examines how avatar representation shapes the perceived quality of gesture animation. Following our first experiment, all stimuli used blurred faces to minimize facial confounds. As shown in Table 1, we added a mismatched mocap condition to ensure the isolation of the confounding effect of motion realism from speech alignment, ensuring that participant evaluations measure the gesture-speech fit rather than the visual quality of the movement.

To create the mismatch mocap condition, we followed GENEA Challenge procedures [Kucherenko et al. 2023]. Specifically, we selected a motion segment from the same actor that began at a different conversational phrase onset and then trimmed it to match the target speech duration, preserving a natural gesture initiation and avoiding visible discontinuities at the start of the clip. All segments were manually checked to ensure the actor remained an "active speaker" and that the animation was free of artifacts, so any perceived degradation could be attributed to the absence of semantic and temporal synchronization. We hypothesized that: **H2A:** Synthetic motion will be rated as less natural on human-detailed avatars than on simpler avatars. **H2B:** Human-detailed avatars will amplify perceived speech–gesture misalignment in the Synthetic and Mismatched conditions relative to simpler avatars. **H2C:** Comprehension will be higher for Matched Mocap than for Mismatched or Synthetic motion, and higher for human-detailed than for simpler avatars. **H2D:** Imperfect motion (Synthetic/Mismatched) will reduce Appeal and increase Eeriness more for human-detailed avatars than for simpler avatars.

5.1 Measurements

To assess motion quality, we adopted two established measures from the GENEA Challenge: motion human-likeness and speech-gesture match [Kucherenko et al. 2023]. Although GENEA mutes audio for human-likeness, we retained speech in all trials to support other perceptual ratings, and instructed participants to ignore speech when rating human-likeness. In addition, perceived comprehension was assessed via a single item measuring gesture-supported understanding, adapted by prior work [Adkins et al. 2023] from the Networked Minds questionnaire [Harms and Biocca 2004]. Finally, to capture affective responses potentially shaped by interactions between appearance and motion cues, including effects related to the uncanny valley, we included Appeal and Eeriness using the same scales as [McDonnell et al. 2012]. All items were rated after each trial using 7-point Likert scales.

Table 2. Summary of statistical main effects and interactions. Measures marked with * were analyzed using Aligned Rank Transform (ART)

Factor	*F*-statistics	*p-value*	η_p^2
Face Representation Experiment			
Speech-Gesture Match			
Motion	$F(1, 127) = 335.26$	$p < .001$	.725
Face	$F(3, 127) = 2.80$	$p < .043$	.062
Anthropomorphism*			
Motion	$F(1, 127) = 172.61$	$p < .001$	.576
Face	$F(3, 127) = 4.38$	$p < .006$	.094
Likeability			
Motion	$F(1, 127) = 101.62$	$p < .001$	.445
Distraction*			
Face	$F(3, 127) = 5.98$	$p < .001$	.124
Avatar Representation Experiment			
Motion Human-likeness*			
Motion	$F(2, 346) = 216.50$	$p < .001$	.556
Avatar	$F(6, 173) = 2.22$	$p < .044$	.071
Avatar ×Motion	$F(12, 346) = 4.55$	$p < .001$	.136
Speech-Gesture Match*			
Motion	$F(2, 346) = 331.81$	$p < .001$	.657
Avatar ×Motion	$F(12, 346) = 2.58$	$p < .003$	.082
Comprehension			
Motion	$F(1.87, 322.94) = 365.14$	$p < .001$	.679
Avatar	$F(6, 173) = 3.00$	$p < .009$	.094
Avatar ×Motion	$F(11.20, 322.94) = 3.03$	$p < .001$	.095
Appeal			
Motion	$F(1.86, 320.98) = 168.62$	$p < .001$	.494
Avatar	$F(6, 173) = 2.79$	$p < .014$	.088
Eeriness*			
Motion	$F(2, 346) = 83.04$	$p < .001$	.324
Avatar	$F(6, 173) = 2.59$	$p < .020$	.083
Avatar ×Motion	$F(12, 346) = 3.29$	$p < .001$	.102
Speech-Muted Validation Experiment			
Mocap-Advantage Score			
Avatar	$F(2, 161) = 17.11$	$p < .001$	.175

5.2 Participants & Procedure

Following the same recruitment criteria as in Experiment 1, we recruited 180 participants (95 female, 85 male; ages 23–58) and distributed them evenly across the seven body representation conditions (approx. 26 per condition). Participants were randomly assigned to one avatar type and completed 9 trials (3 Motion Conditions × 3 Scenarios) presented in randomized order. AI model assignment followed the same procedure as in Experiment 1.

5.3 Results

Experiment 2 followed the same statistical protocol. As the within-subject factor *Motion* had three levels, sphericity was assessed using Mauchly's test; when violated, Greenhouse–Geisser corrections were applied. Effect sizes (Cohen's d) are additionally reported for

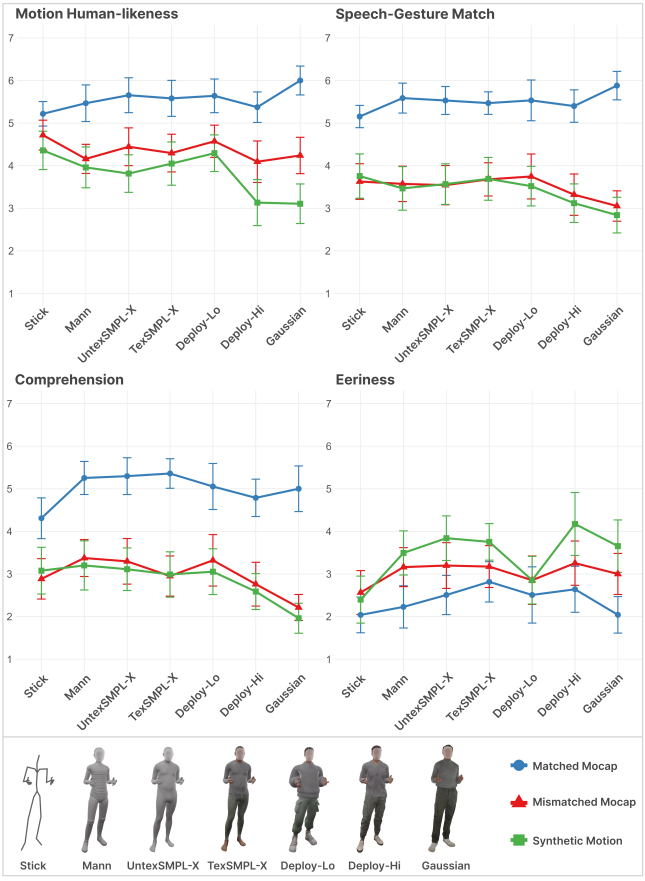


Fig. 4. Interaction effect between Avatar and Motion on various scales in Avatar Representation Experiment.

within-avatar simple effects to quantify motion condition discriminability (see Fig.4).

Motion Human-likeness. Motion Condition strongly affected human-likeness: matched mocap was rated most natural, followed by mismatched mocap, with synthetic motion lowest (all $p < .001$). Avatar Type also had an effect, with the Deploy-Lo rated more humanlike than the Deploy-Hi avatar ($p < .047$). Importantly, these effects interacted: only the Gaussian avatar showed clear separation across all three motion levels (Matched > Mismatched > Synthetic, all $p < .05$), as shown in Fig.4. For the other avatars (Deploy-Hi, TexSMPL-X, UntexSMPL-X, Mann, and Deploy-Lo), matched mocap was better than mismatched and synthetic (all $p < .05$), but mismatched and synthetic did not differ. The stick figure showed the weakest separation (Matched > Synthetic only, $p < .015$), with Matched–Synthetic effect sizes reducing from $d = 2.65$ (Gaussian) to $d = 0.88$ (stick figure).

Speech-Gesture Match. A large main effect of Motion Condition was observed. Matched mocap was rated significantly higher than both synthetic and mismatched motion (both $p < .001$), which did not differ from one another. While Avatar Type yielded no main effect, there was a significant Avatar Type × Motion Condition interaction (see Fig.4). Specifically, the preference for Matched motion held across all avatars (all $p < .001$), but the effect was most pronounced for the Gaussian avatar (Matched Mocap–Mismatched Mocap $d = 2.33$; Matched Mocap–Synthetic $d = 2.45$) and least pronounced for the Stick figure (Matched Mocap–Mismatched Mocap $d = 1.31$; Matched Mocap–Synthetic $d = 1.17$). Other avatars fell within the $d = 1.40$ to 1.83 range.

Comprehension. A significant main effect of Motion Condition revealed higher perceived comprehension for Matched Mocap compared to both Synthetic and Mismatched Mocap (all $p < .001$), with no difference between Synthetic and Mismatched. Avatar Type also had a main effect: participants perceived Gaussian as lower in comprehension than Mann ($p = .028$) and UntexSMPL-X ($p = .038$). These effects were qualified by a significant Avatar × Motion interaction (see Fig.4). While the drop from Matched to Synthetic motion was significant for all avatars, the magnitude of this effect was most pronounced for the Gaussian avatar ($d = 1.81$), whereas the Stick Figure was notably more robust to the use of synthetic motion ($d = 0.75$). Furthermore, in the Mismatched and Synthetic conditions, the Gaussian avatar was consistently rated lower for comprehension than the Deploy-Lo character, Human Mannequin, and Untextured SMPL-X models ($p < .05$).

Appeal. Significant main effects were found for both Motion and Avatar. Matched Mocap was rated significantly higher than Mismatched ($p < .001$), which in turn was rated higher than Synthetic ($p < .001$). Across body types, the Deploy-Lo character was rated significantly more appealing than the Deploy-Hi avatar ($p = .044$).

Eeriness. Results showed significant main effects: Deploy-Hi avatar was eerier than stick figure ($p < .032$); Matched mocap was less eerie than Mismatched ($p < .001$) and Synthetic ($p < .001$), with Synthetic eerier than Mismatched ($p < .001$). A significant Body Representation × Motion interaction was found (see Fig.4). Post-hoc comparisons revealed a Synthetic > Matched Mocap eeriness effect for Gaussian, Deploy-Hi, Mannequin, and both mesh avatars (all $p < .01$), but not for the Deploy-Lo or Stick Figure avatars. Additionally, Mismatched Mocap was more eerie than Matched Mocap (all $p < .01$) on Gaussian and Mann avatars. Under synthetic motion, the Stick Figure was significantly less eerie than the Deploy-Hi and mesh avatars.

5.4 Experiment 2 Discussion

This experiment produced two core findings. First, the motion manipulation worked as intended: Matched Mocap was rated best overall relative to Synthetic and Mismatched Mocap. Second, Gaussian avatar made motion artifacts and the eeriness of speech–gesture mismatches much easier to notice. In contrast, stick figures often compressed differences.

However, an important confound for interpreting motion human-likeness was also revealed: although Matched and Mismatched conditions both used ground truth mocap, Mismatched Mocap was rated less human-like. This suggests that participants penalized perceived human-likeness when gestures and speech were misaligned even

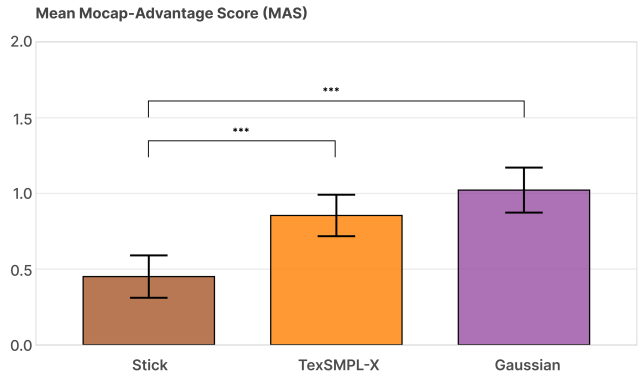


Fig. 5. Mocap-Advantage Score (MAS) by Avatar Type in Speech-Muted Validation.

when instructed to ignore speech. To isolate motion quality sensitivity from alignment effects, we therefore conducted Experiment 3 using speech-muted, pairwise comparisons.

6 Experiment 3: Speech-Muted Validation

As shown in Table 1, we selected 3 representative avatars, which mirror the strongest to weakest motion quality separation observed in Experiment 2. Each participant was assigned to one avatar condition and evaluated motion sequences derived from two sources: Mocap and synthetic motion. We adopted a pairwise comparison protocol similar to the GENEA challenges [Nagy et al. 2025]. Participants performed a side-by-side comparison task where two silent videos (Mocap vs. Synthetic) were presented simultaneously for the same underlying sequence.

We quantified discriminability with a Mocap-Advantage Score (MAS). On each trial, participants chose the more natural video and rated preference strength on a 5-point scale (strong/weak for either side, or tie). Responses were coded so positive values favored mocap and negative values favored synthetic. MAS was computed as the mean coded response across nine trials per participant; higher MAS indicates a stronger mocap advantage (greater sensitivity to motion-quality differences) under the assigned avatar.

6.1 Participants & Procedure

We recruited 164 participants via Prolific (77 female, 87 male; ages 18–55), excluding 16 for failed attention checks. Participants were evenly distributed across the three avatar conditions (approx. 55 per condition).

6.2 Results

After confirming that assumptions were satisfied (as in the previous analysis), a one-way ANOVA showed a significant effect of avatar on the outcome variable. Post-hoc comparisons showed that Gaussian and TexSMPL-X yielded higher discriminability than Stick (All $p < .001$), with a larger standardized difference for Gaussian relative to Stick ($d = 1.09$) than for TexSMPL-X relative to Stick ($d = 0.77$). Gaussian and TexSMPL-X did not differ significantly (Fig. 5).

7 General Discussion

Our Experiment 1 offers a clear methodological insight for the community: in motion-generation studies, facial presentation can systematically bias judgments of motion quality. Many recent papers present face and body motion together [Liu et al. 2024], yet our results suggest that even minimal facial animation can inflate ratings on core gesture outcomes by boosting overall audiovisual coherence. This creates a risk of misattribution, where evaluators credit the gesture itself for integration cues that are actually driven by the face. Conversely, a common control choice is to use a static face (e.g., [Cheng and Huang 2025]); however, a frozen face on a moving body can heighten cue mismatch and reduce perceived human-likeness. Face occlusion with an opaque mask (as in [Nagy et al. 2025]) was suboptimal in our setting, as it was more distracting than other approaches and it may have diverted attention from the evaluated motion. By contrast, facial blurring provided the best trade-off: it minimized distraction, avoided frozen-face artifacts, and reduced evaluation bias in speech–gesture alignment. *We therefore recommend controlling facial information by blurring as a stable, low-bias evaluation baseline when isolating gesture quality.*

Experiments 2 and 3 tested how motion quality is perceived across avatar representations. Human-captured motion with matched speech received the highest human-likeness ratings, consistent with the intended motion-quality manipulation. Crucially, motion-condition separability depended on avatar type: the Gaussian avatar most clearly distinguished the three motion conditions, followed by DeployHi, while the Stick compressed differences and often rendered Mismatched and Synthetic indistinguishable. This pattern supports H2A and aligns with prior work showing that reduced or featureless body representations mask kinematic artifacts [Chaminade et al. 2007].

A potential confound is that Mismatched Mocap, despite using ground-truth motion capture, was rated as less human-like, suggesting that participants incorporated speech–gesture coherence into human-likeness judgments even when instructed to ignore speech. This aligns with previous findings that impressions of motion quality and cross-modal alignment are not fully separable [Nagy et al. 2025]. To isolate kinematic quality from alignment, we conducted a speech-muted validation. Discriminability remained higher for Gaussian and TexSMPL-X than for Stick, indicating that motion human-likeness differences are most clearly differentiated on the Gaussian avatar even when speech cues are removed.

For speech gesture match, participants preferred human captured motion with matched speech, but rated Mismatched and Synthetic versions equally. This indicates that, at the synthetic quality tested here, generated gestures did not outperform a deliberately mismatched mocap baseline. Avatar type had no main effect on match but moderated the Motion effect: the Matched–Mismatched gap was largest for the Gaussian avatar and smallest for the stick figure, supporting H2B. Importantly, because the Mismatched condition retained the same mocap motion, this gap reflects a penalty for semantic and temporal misalignment rather than reduced motion human-likeness. Whereas prior work often varies perceived human-likeness via keyframe interpolation or mocap–synthetic blends [Chaminade et al. 2007; Ferstl et al. 2021], our findings are the first to show that even ground-truth mocap is penalized when

gestures are temporally or semantically incongruent with speech. *Overall, Gaussian avatars consistently yielded the clearest separation between motion conditions for both motion human-likeness and speech gesture match, compared to the other body representations.*

Comprehension results indicate that only Matched Mocap improved comprehension; Mismatched Mocap (despite being mocap) was indistinguishable from Synthetic, suggesting that gesture–speech alignment, rather than motion human-likeness alone, drives comprehension benefits. Contrary to H2C, avatar appearance interacts with motion condition: when transitioning from Matched to Synthetic motion, Gaussian exhibits the largest decline in comprehension, and under Synthetic motion it becomes the hardest avatar to interpret, performing worse than the standard research baselines (SMPL-X and Mann) as well as the Deploy-Lo avatar. Here, the Gaussian's complex visual style, combined with motion that is incongruent with its appearance, may introduce additional processing overhead [Kätsyri et al. 2015], reducing intelligibility. *These findings suggest that Gaussian is a poor fit when motion is unreliable, whereas a production-ready avatar such as Deploy-Lo is suitable for deployment and preserves comprehension for synthetic motion.*

Beyond communicative efficiency, motion and appearance style strongly shaped affect. Across appeal and eeriness, responses were driven more by motion's biological plausibility than semantic alignment: appeal declined from Matched Mocap to Mismatched Mocap to Synthetic, while eeriness rose in reverse (Matched least eerie, Synthetic most eerie). This suggests biologically fluid motion is inherently more pleasing than unnatural artifacts, even when gestures mismatch speech [Amadou et al. 2023; Ferstl et al. 2021], consistent with heightened sensitivity to non-biological motion and increased processing demands [Saygin et al. 2012].

Appearance factors further explained when motion flaws become negative. For appeal, Deploy-Lo was rated as more appealing than Deploy-Hi, likely because its exaggerated outfit proportions read as more stylized, reducing strict human realism expectations [Kätsyri et al. 2015]. For eeriness, Body Representation moderated the effect of motion condition: avatars with more humanlike form (Gaussian, Deploy-Hi, textured/untextured SMPL-X, and Mann) appeared to require high motion integrity to avoid eeriness, whereas the Stick remained comparatively low in eeriness across conditions. Deploy-Lo showed a similar robustness profile to the Stick, consistent with its higher appeal. Overall, this pattern reinforces appearance–motion compatibility as a core driver of uncanny impressions [Rekik et al. 2024]. *Overall, Stick and Deploy-Lo are affectively more forgiving, whereas Gaussian avatar with richer human detail demands higher motion integrity, otherwise appeal drops and eeriness rises sharply.*

8 Limitations & Future Work

Despite these findings, several important considerations should be noted. First, our results are specific to the stimuli tested; future work should cover a broader range of communicative intents and emotions. We also limited avatars to human-like proportions to avoid retargeting issues common with exaggerated characters, and all stimuli used a single male Gaussian avatar whose subject was reconstructed from the dataset in [Işık et al. 2023], with clothing kept as consistent as possible. Since the dataset lacked a female avatar with closely matched clothing, adding one would have confounded gender with outfit differences; future work should include greater variation in body shape and gender diversity. Additionally, the facial animation in this study was limited; future work should increase facial animation realism to test whether it linearly inflates gesture-related ratings. Our Gaussian avatar is point-based, with motion introduced via animation, rather than explicit temporal-dimension modeling of the avatar [Luiten et al. 2024]; incorporating more advanced avatars in future work would further validate our results. We pooled multiple generative models to estimate an overall synthetic-gesture effect; model-level comparisons will require adequately powered studies. Separately, we made a few design choices and outline broader next steps. Given that clip duration affects perception of motion artifacts [Adkins et al. 2023], we used 20s stimuli (rather than GENEA's 10s clips [Nagy et al. 2025]) to better approximate real interactions and reduce chance-driven judgments. Future work could further explore VR contexts [Du et al. 2025], add objective measures (e.g., eye-tracking, forced-choice discrimination), and extend evaluation to dyadic interactions, especially given recent advances in synthesizing high-fidelity listening behavior [Agrawal et al. 2025].

9 Conclusions

Across three experiments, we provide the first systematic evidence that a virtual character's visual form influences how people judge generated gesture animations. Meanwhile, speech-driven gesture generation is improving rapidly, driven primarily by large-scale motion-capture datasets that enable training higher-capacity models. As synthetic motion nears captured ground truth, a controlled perceptual evaluation is increasingly necessary to detect subtle artifacts and compare methods fairly. Our results offer timely guidance for benchmarking and for deploying generated motion.

Acknowledgments

We would like to thank Kiran Chhatre for helpful discussions on avatar and gesture generation model selection. This work was conducted with the financial support of the Research Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224.

References

Alex Adkins, Aline Normoyle, Lorraine Lin, Yu Sun, Yuting Ye, Massimiliano Di Luca, and Sophie Jörg. 2023. How important are detailed hand motions for communication for a virtual character through the lens of charades? *ACM Transactions on Graphics* 42, 3 (2023), 1–16.

Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, et al. 2025. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. *arXiv preprint arXiv:2506.22554* (2025).

Ghazanfar Ali, Woojoo Kim, Muhammad Shahid Anwar, Jae-In Hwang, and Ahyoung Choi. 2025. Expanding Multilingual Co-Speech Interaction: The Impact of Enhanced Gesture Units in Text-to-Gesture Synthesis for Digital Humans. *IEEE Access* (2025).

Nabila Amadou, Kazi Injamamul Haque, and Zerrin Yumak. 2023. Effect of appearance and animation realism on the perception of emotionally expressive virtual humans. In *Proceedings of the 23rd ACM international conference on intelligent virtual agents*. 1–8.

Christoph Bartneck, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. My robotic doppelgänger-A critical look at the uncanny valley. In *RO-MAN 2009-The 18th IEEE international symposium on robot and human interactive communication*. IEEE, 269–276.

Thierry Chaminade, Jessica Hodgins, and Mitsuo Kawato. 2007. Anthropomorphism influences perception of computer-animated characters' actions. *Social cognitive and affective neuroscience* 2, 3 (2007), 206–216.

Bohong Chen, Yumeng Li, Youyi Zheng, Yao-Xiang Ding, and Kun Zhou. 2025. Motion-example-controlled Co-speech Gesture Generation Leveraging Large Language Models. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. 1–12.

Hongye Cheng, Tianyu Wang, Guangsi Shi, Zexing Zhao, and Yanwei Fu. 2025. HOP: Heterogeneous Topology-based Multimodal Entanglement for Co-Speech Gesture Generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 906–916.

Yongkang Cheng and Shaoli Huang. 2025. HoloGest: Decoupled Diffusion and Motion Priors for Generating Holistically Expressive Co-Speech Gestures. In *Proceedings of the International Conference on 3D Vision (3DV)*.

Kiran Chhatre, Nikos Athanasiou, Giorgio Becherini, Christopher Peters, Michael J Black, Timo Bolkart, et al. 2024. Emotional speech-driven 3d body animation via disentangled latent diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1942–1953.

Sharice Clough and Melissa C Duff. 2020. The role of gesture in communication and cognition: Implications for understanding and treating neurogenic communication disorders. *Frontiers in Human Neuroscience* 14 (2020), 323.

Karlo Crnek, Grega Močnik, and Matej Rojc. 2025. Advancing objective evaluation of speech-driven gesture generation for embodied conversational agents. *International Journal of Human–Computer Interaction* (2025), 1–17.

Lipeng Cui and Jiarui Liu. 2023. Virtual human: A comprehensive survey on academic and applications. *IEEE Access* 11 (2023), 123830–123845.

Michael J Doughty. 2001. Consideration of three types of spontaneous eyeblink activity in normal humans: during reading and video display terminal use, in primary gaze, and while in conversation. *Optometry and vision science* 78, 10 (2001), 712–725.

Haoyang Du, Kiran Chhatre, Christopher Peters, Brian Keegan, Rachel McDonnell, and Cathy Ennis. 2025. Synthetically Expressive: Evaluating gesture and voice for emotion and empathy in VR and 2D scenarios. In *Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*. 1–10.

Ylva Ferstl, Sean Thomas, Cédric Guiard, Cathy Ennis, and Rachel McDonnell. 2021. 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*. 76–83.

Kazi Injamamul Haque, Alkiviadis Pavlou, and Zerrin Yumak. 2025. "Wild West" of Evaluating Speech-Driven 3D Facial Animation Synthesis: A Benchmark Study. In *Computer Graphics Forum*. Wiley Online Library, e70073.

Chad Harms and Frank Biocca. 2004. Internal consistency and reliability of the networked minds measure of social presence. (2004).

Roy S Hessels. 2020. How does gaze to faces support face-to-face interaction? A review and perspective. *Psychonomic bulletin & review* 27, 5 (2020), 856–881.

Darragh Higgins, Katja Zibrek, Joao Cabral, Donal Egan, and Rachel McDonnell. 2022. Sympathy for the digital: Influence of synthetic voice on affinity, social presence and empathy for photorealistic virtual humans. *Computers & Graphics* 104 (2022), 116–128.

Chin-Chang Ho and Karl F MacDorman. 2010. Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. *Computers in Human Behavior* 26, 6 (2010), 1508–1518.

Judith Holler, Louise Schubotz, Spencer Kelly, Peter Hagoort, Manuela Schuetze, and Aslı Özyürek. 2014. Social eye gaze modulates processing of speech and co-speech gesture. *Cognition* 133, 3 (2014), 692–697.

Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. 2023. HumanRF: High-Fidelity Neural Radiance Fields for Humans in Motion. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–12. doi:10.1145/3592415

Jari Kätsyri, Klaus Förger, Meeri Mäkäräinen, and Tapio Takala. 2015. A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in psychology* 6 (2015), 390.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.* 42, 4, Article 139 (July 2023), 14 pages. doi:10.1145/3592433

Anna Krason, Rebecca Fenton, Rosemary Varley, and Gabriella Vigliocco. 2022. The role of iconic gestures and mouth movements in face-to-face communication. *Psychonomic bulletin & review* 29, 2 (2022), 600–612.

Taras Kucherenko, Rajmund Nagy, Youngwoo Yoon, Jieyeon Woo, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2023. 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*. 792–801.

Taras Kucherenko, Pieter Wolfert, Youngwoo Yoon, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2024. Evaluating gesture generation in a large-scale open challenge: The GENEA Challenge 2022. *ACM Transactions on Graphics* 43, 3 (2024), 1–28.

Peter Kullmann, Timo Menzel, Mario Botsch, and Marc Erich Latoschik. 2023. An evaluation of other-avatar facial animation methods for social VR. In *Extended abstracts of the 2023 CHI conference on human factors in computing systems*. 1–7.

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. 2024. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1144–1154.

Pinxin Liu, Luchuan Song, Junhua Huang, Haiyang Liu, and Chenliang Xu. 2025. GestureLSM: Latent Shortcut based Co-Speech Gesture Generation with Spatial-Temporal Modeling. *arXiv preprint arXiv:2501.18898* (2025).

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. 2024. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision (3DV)*. IEEE, 800–809.

Karl F MacDorman and Debaleena Chattopadhyay. 2016. Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition* 146 (2016), 190–205.

Rachel McDonnell, Martin Breidt, and Heinrich H Bülthoff. 2012. Render me real? Investigating the effect of render style on the perception of animated virtual humans. *ACM Transactions on Graphics (TOG)* 31, 4 (2012), 1–11.

Rachel McDonnell, Sophie Jörg, Joanna McHugh, Fiona Newell, and Carol O'Sullivan. 2008. Evaluating the emotional content of human motions on real and virtual characters. In *Proceedings of the 5th symposium on Applied perception in graphics and visualization*. 67–74.

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

Rajmund Nagy, Hendric Voss, Thanh Hoang-Minh, Mihail Tsakov, Teodor Nikolov, Zeyi Zhang, Tenglong Ao, Sicheng Yang, Shaoli Huang, Yongkang Cheng, et al. 2025. Gesture Generation (Still) Needs Improved Human Evaluation Practices: Insights from a Community-Driven State-of-the-Art Benchmark. *arXiv preprint arXiv:2511.01233* (2025).

Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. 2024. From audio to photoreal embodiment: Synthesizing humans in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1001–1010.

Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. 2023. A comprehensive review of data-driven co-speech gesture generation. In *Computer Graphics Forum*, Vol. 42. Wiley Online Library, 569–596.

Catherine Oh Kruzic, David Kruzic, Fernanda Herrera, and Jeremy Bailenson. 2020. Facial expressions contribute more than body movements to conversational outcomes in avatar-mediated virtual environments. *Scientific reports* 10, 1 (2020), 20626.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10975–10985.

Xingqun Qi, Jiahao Pan, Peng Li, Ruibin Yuan, Xiaowei Chi, Mengfei Li, Wenhan Luo, Wei Xue, Shanghang Zhang, Qifeng Liu, et al. 2024. Weakly-supervised emotion transition learning for diverse 3d co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10424–10434.

Rim Rekik, Stefanie Wuhrer, Ludovic Hoyet, Katja Zibrek, and Anne-Hélène Olivier. 2024. A survey on realistic virtual human animations: Definitions, features and evaluations. In *Computer Graphics Forum*, Vol. 43. Wiley Online Library, e15064.

Ayse Pinar Saygin, Thierry Chaminade, Hiroshi Ishiguro, Jon Driver, and Chris Frith. 2012. The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social cognitive and affective neuroscience* 7, 4 (2012), 413–422.

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.

Leslie Wöhler, Satoshi Ikehata, and Kiyoharu Aizawa. 2024. Investigating the Perception of Facial Anonymization Techniques in 360° Videos. *ACM Transactions on Applied Perception* 21, 4 (2024), 1–17.

Pieter Wolfert, Nicole Robinson, and Tony Belpaeme. 2022. A review of evaluation practices of gesture generation in embodied conversational agents. *IEEE Transactions on Human-Machine Systems* 52, 3 (2022), 379–389.

Yinghan Xu, Théo Morales, and John Dingliana. 2026. SplatBus: A Gaussian Splatting Viewer Framework via GPU Interprocess Communication. *arXiv preprint arXiv:2601.15431* (2026).

Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. 2023. Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. *arXiv preprint arXiv:2305.04919* (2023).

Eduard Zell, Carlos Aliaga, Adrian Jarabo, Katja Zibrek, Diego Gutierrez, Rachel McDonnell, and Mario Botsch. 2015. To stylize or not to stylize? The effect of shape and material stylization on the perception of computer-generated faces. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 1–12.

Youyi Zhan, Tianjia Shao, Yin Yang, and Kun Zhou. 2025. Real-time High-fidelity Gaussian Human Avatars with Position-based Interpolation of Spatially Distributed MLPs . In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 26297–26307. doi:10.1109/CVPR52734.2025.02449

Zeyi Zhang, Tenglong Ao, Yuyao Zhang, Qingzhe Gao, Chuan Lin, Baoquan Chen, and Libin Liu. 2024. Semantic gesticulator: Semantics-aware co-speech gesture synthesis. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–17.

Katja Zibrek and Rachel McDonnell. 2014. Does render style affect perception of personality in virtual humans?. In *Proceedings of the ACM symposium on applied perception*. 111–115.