Synthetically Expressive: Evaluating gesture and voice for emotion and empathy in VR and 2D scenarios

Haoyang Du
Technological University Dublin
Dublin, Ireland
D23128268@mytudublin.ie

Kiran Chhatre
KTH Royal Institute of Technology
Stockholm, Sweden
chhatre@kth.se

Christopher Peters
KTH Royal Institute of Technology
Stockholm, Sweden
chpeters@kth.se

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
Dublin, Ireland
cathy.ennis@mu.ie



Figure 1: Participants in 2D screen (bottom left corner) VR (bottom right corner) immersion conditions. The virtual human delivers a one-minute monologue in three emotional scenarios (positive, neutral, negative) in random order, with varying gesture and voice realism. Participants then provide ratings on their experience.

Abstract

The creation of virtual humans increasingly leverages automated synthesis of speech and gestures, enabling expressive, adaptable agents that effectively engage users. However, the independent development of voice and gesture generation technologies, alongside the growing popularity of virtual reality (VR), presents significant questions about the integration of these signals and their ability to convey emotional detail in immersive environments. In this paper, we evaluate the influence of real and synthetic gestures and speech, alongside varying levels of immersion (VR vs. 2D displays) and emotional contexts (positive, neutral, negative) on user perceptions. We investigate how immersion affects the perceived match between gestures and speech and the impact on key aspects of user experience, including emotional and empathetic responses and the sense of co-presence. Our findings indicate that while VR enhances the perception of natural gesture–voice pairings, it does not similarly improve synthetic ones—amplifying the perceptual gap between them. These results highlight the need to reassess gesture appropriateness and refine AI-driven synthesis for immersive environments.

CCS Concepts

• **Computing methodologies** → *Intelligent agents*; *Motion capture*; **Virtual reality**; **Perception**.

Keywords

ML for animation, Speech synthesis, User studies, Perception in VR

ACM Reference Format:
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 *ACM International Conference on Intelligent Virtual Agents (IVA '25), September 16–19, 2025, Berlin, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3717511.3747074

This work is licensed under a Creative Commons Attribution 4.0 International License.
IVA '25, Berlin, Germany
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1508-2/25/09
https://doi.org/10.1145/3717511.3747074

1 Introduction

Advances in Artificial Intelligence (AI) and Virtual Reality (VR) technologies have profoundly reshaped computer-mediated interactions. Virtual humans — computer-generated avatars designed to

engage users — are now widely utilized in customer service [83], education [17] and healthcare [65]. Their effectiveness relies on the seamless integration of verbal and non-verbal communication channels — gestures and speech — to build rapport, convey emotions and evoke empathetic responses [19, 63, 78]. Human communication is inherently multimodal, therefore, accurately modelling these behaviours is vital for creating believable virtual agents [80].

Although AI-driven co-speech gesture synthesis and modern text-to-speech (TTS) systems have become increasingly human-like, many still lack the adaptability needed to replicate diverse emotional expressions [54, 77]. In less immersive interfaces, enhancing a single modality, such as gestures or speech, improves acceptance and trustworthiness, supporting the "maximize" hypothesis that advocates for the highest-quality output in each channel [31, 57].

However, VR's elevated immersiveness amplifies users' sensitivity to mismatched signals [46], owing to its potent illusions of place [71] and social presence [55]. Consistency in motion and appearance is more critical in immersive environments than in less immersive 2D interfaces, emphasizing the importance of coherent multimodal cues [52]. Emotional and empathic responses, which are central to meaningful interactions [58], are undermined by mismatches between verbal and non-verbal empathic behaviors [57]. Empirical evidence further highlights that congruence across channels significantly impacts emotional engagement [14, 38]. These findings emphasize the importance of advancing text-to-speech and gesture integration for virtual characters while also examining how immersion affects users' sensitivity to realism.

To address this gap in understanding how immersion impacts users' sensitivity to mismatched signals, this paper investigates the influence of different gesture-voice combinations on perceptions of virtual human interactions in both 2D and immersive VR environments. We conducted an experiment where participants watched a one-minute monologue delivered by a virtual character. The character's gestures and voice varied in realism, and were presented across different emotional scenarios in either a 2D or VR context. We find that immersive VR environments widen the perceptual gap between natural and synthetic gestures, as natural gestures become more salient while synthetic ones do not receive the same enhancement. Notably, the presence of even one natural cue—either gesture or voice—elicits emotional and empathic responses comparable to fully natural conditions, underscoring the value of partial naturalism in immersive settings.

2 Related Work

Expressive Gesture Generation: Virtual humans often employ co-speech gestures mirroring real-human behaviours, essential for nonverbal communication to help convey ideas, and express emotions [43]. Recent progress in datasets [21, 32, 42, 48, 49] has produced deep learning–based gesture generation from audio [47] text [9], rhythm[4] and contextual cues [8, 79]. Current models are increasingly designed to learn multiple modalities or features to generate more natural and expressive gestures [5, 33, 56, 86], including gesture style transfer methods that aim to alter the perceived impression of gestures while preserving their semantic and temporal structure [85]. Among recent methods, emotion-conditioned gesture synthesis has received significant attention: one approach encodes emotion labels within a temporal convolutional network [49], while another aligns gestures with emotional beats but struggles with rare emotions [62]. A novel line of work disentangles speech into latent vectors capturing content, style, and emotion for more flexible control [13]. These strategies may struggle with subtle expressions and require extensive data for smooth results [54], highlighting the need to evaluate users' perception of synthesized gestures' emotional intent in interactive contexts.

Subjective evaluation of synthetic gestures often focuses on naturalness and appropriateness through user studies in 2D environments [45], but it lacks consideration of emotional expressiveness. Gestures play a crucial role in conveying emotions, even in the absence of facial expressions [28]. Well-timed, spatially coherent gestures can enhance users' sense of presence, emotional engagement, and realism [29]. For instance, virtual characters with high-fidelity animations have been shown to boost social presence and evoke stronger emotional responses [82]. Real-time interaction with data-driven gestures can also enhance user engagement [35]. However, exaggerated movements combined with unsettling appearances, such as zombie-like features, can increase arousal and intensify emotions while reducing likeability, emphasizing the need to balance realism and exaggeration [53]. Additionally, VR has become a powerful tool for emotional induction in psychological studies [72], highlighting the importance of animation accuracy for social interactions in VR [64]. Additional research is needed to understand the emotional expressiveness of synthetic gestures in virtual environments, particularly the differences between 2D and VR contexts.

Expressive Text-To-Speech: Text-to-speech (TTS) technology facilitates the conversion of written text into spoken words, with the objective of generating speech that closely approximates natural human communication [75]. Early TTS systems focused on intelligibility using articulatory, formant, and concatenative synthesis but often produced unnatural, mechanical speech, reducing user engagement [11]. Recent advancements in Expressive Text-to-Speech (E-TTS), as detailed by Kaur and Singh [41], represent a significant progression in this domain. E-TTS incorporates emotional expressiveness, speaking styles, and contextual vocal adjustments to enhance the naturalness and user experience of synthetic speech [6]. This integration results in synthesized speech that is nearly indistinguishable from human voices. The perception of advanced neural TTS systems remains a key focus in human-computer interaction research. TTS voices have been found to be less humanlike and less trustworthy than real voices [15]. Studies comparing human voices with advanced neural TTS—using deep neural networks for natural speech—suggest that human voices are rated as more trustworthy, even though no perceptual differences in voice quality were identified [23]. In contrast, in VR, TTS voices are rated less appealing, expressive, and compatible with a character's appearance, reducing social presence and emotional responses [38]. These findings highlight the need for further research to examine the perceptual differences of TTS voices in 2D versus VR environments.

Mixed Realism Effects: Inconsistent levels of realism in virtual human features can lead to negative impressions, commonly referred to as the uncanny valley effect [50]. One study found that mismatched voice and appearance affected how anthropomorphic

the agent seemed to participants [2]. Variations in appearance and motion have been shown to influence emotional perception [3, 53]. Additionally, insufficient emotional expressiveness further contributes to this phenomenon [76]. Researchers have explored how virtual humans expressing emotions beyond neutrality influence user perceptions, finding that high-fidelity characters displaying positive emotions are perceived as happier than low-fidelity ones, with a less pronounced effect for negative emotions [36]. Moreover, subtle asynchronies between verbal and non-verbal cues can disrupt emotional interpretation [30]. Recent findings further indicate that aligning high-fidelity facial motion with realistic appearance significantly enhances users' empathy across a range of emotional scenarios [37, 38]. These findings suggest that high realism cues may enhance emotional engagement in positive scenarios, highlighting the need to examine whether gestures shape empathy when TTS and appearance realism lack impact.

Some research supports the maximization hypothesis, i.e., that each communicative channel (e.g., animation, voice) should prioritize high realism rather than maintaining uniform fidelity across all features [31, 57]. These studies identified TTS voice paired with robotic animation as the most mismatched combination, leading to notably low user preference [31]. Immersion further influences virtual human perception, with VR studies revealing dynamics distinct from those observed in 2D displays. While 2D research aligns with the maximization hypothesis, VR heightens sensitivity to feature congruence, enabling congruent virtual humans to appear more plausible and less eerie [52]. We anticipate that VR will enhance the impact of human gestures and speech, amplifying their perceptual distinction from AI-based gesture-voice combinations.

Co-Presence: Co-presence, the feeling of sharing a virtual space with others, is fundamental to effective VR interactions [70]. While visual realism contributes to a sense of presence, behavioural interactions play a significant role in enhancing co-presence [55]. Studies indicate that synchronized behaviour is more impactful than visual fidelity on social presence [3], emphasizing the importance of well-timed and realistic behavioural cues in creating engaging experiences [16]. Additionally, voice realism significantly influences co-presence in VR environments. TTS voices often diminish social presence in VR, but have minimal impact in less immersive 2D settings, suggesting that VR amplifies the negative effects of TTS voices [38]. These findings highlight the need to balance visual, behavioural, and auditory elements to optimize co-presence in VR environments across varying immersion levels.

3 Study Overview

We conducted a factorial experiment to investigate how varying forms of speech and gesture, coupled with different levels of immersion, influence participants' perceptions of virtual characters. Utilizing a 3 × 4 × 2 mixed factorial design (See Table S1 in the supplementary PDF and the supplementary video for study design details.), our factors included **Scenario** with three levels (Positive, Neutral, and Negative emotional contexts) as a within-subjects factor, **Gesture-Voice Combination (GVC)** with four levels (GnVn: Natural Gesture × Natural Voice; GnVs: Natural Gesture × Synthetic Voice; GsVn: Synthetic Gesture × Natural Voice; GsVs: Synthetic Gesture × Synthetic Voice) as a between-subjects factor and **Immersion** with two levels (Virtual Reality [VR] and 2D Screen-based Display [2D]) as a between-subjects factor. We present the following hypotheses:

H1: Increasingly synthetic voice and gesture combinations will lead to reduced perceptions of gesture-speech match, likeability, anthropomorphism, emotional responses, empathy, and co-presence.

H2: Higher levels of immersion will amplify the effects of H1, with VR enhancing the positive impact of natural voice and speech while further widening the differences for synthetic combinations.

H3: Characters exhibiting more realistic gestures and speech will elicit higher emotional responses in positive scenarios.

3.1 Measurements

We employed multiple validated measures to evaluate participants' responses. Perceived match between gesture and voice was assessed using a 7-point Likert scale, following [31], who adapted it from the GENEA challenge's continuous scale [44]. Likeability and anthropomorphism of characters were measured using items from the Godspeed questionnaire [39]. Co-presence was evaluated using the Networked Minds Social Presence scale [34], focusing specifically on co-presence to capture the perceptual awareness of a virtual character's presence in our non-interactive study. Emotional responses was measured using the Affective Slider [7], for assessing valence and arousal, where participants adjusted sliders to indicate their levels for each dimension. Finally, empathy was assessed using the Measure of State Empathy [60], which includes three self-report items for Cognitive, Affective, and Compassionate Empathy, with scores computed by summing items for each dimension as outlined in [37]. The specific questionnaire items are provided in Table 2 of the supplementary material.

3.2 Participants

We recruited 219 participants (111 female, 106 male, 2 other; aged 18–54, M = 28.9), who experienced either VR or 2D immersion with one of four gesture-voice combinations across three emotion scenarios. The VR group included 105 participants (M = 27.4, SD = 6.79), recruited via university outreach. The 2D group included 114 participants (M = 30.2, SD = 7.02), recruited through Prolific[1], an online platform widely used in gesture research [45, 84]. To ensure consistency, recruitment was limited to participants from the same geographic region, aged between 18 and 40, and holding at least a bachelor's degree. Additional eligibility requirements included a minimum approval rating of $\geq$ 90%, completion of at least 100 prior studies on Prolific. We configured Prolific to prevent duplicate participation to ensure participants only viewed a single between-group condition.

3.3 Experiment Procedures

VR condition participants used a Meta Quest 3 headset to operate a Unity-based standalone application within a controlled laboratory setting. In the 2D condition, participants downloaded and ran the application on their computer, preserving intended quality and avoiding issues like video compression or latency. At the start of the study, participants signed a consent form and provided their

[1] https://www.prolific.com

demographic information. They then entered a virtual room where they received a brief training session on how to navigate the VR or 2D setup. Following the tutorial, participants observed a virtual character — either male or female — speaking to them within the virtual environment. After each scenario, participants completed a questionnaire within the same immersive setting to maintain their sense of immersion [61]. To ensure data quality, a content-based attention check was included for 2D participants only. Those who failed were excluded from analysis. As the VR sessions were supervised in person, no attention check was required. The VR sessions lasted approximately 15–20 minutes due to the additional setup required, including fitting the headset and adjusting it for comfort. To thank them for their time, we held a raffle for two £50 vouchers for VR participants. 2D condition participants completed the task in an average of 12 minutes, as this condition did not require such preparations. For their time, 2D participants received €2 compensation, equivalent to €10 per hour. Ethical approval for the study was obtained from the institutional ethics committee.

3.4 Stimuli Creation

Our virtual characters, developed in Character Creator 4 (CC4)[2], included two genders with consistent appearance to enhance variation and realism in the experiment. In Unity, we used the Standard Render Pipeline for CC4 asset compatibility and integrated CC4 shaders with basic lighting and post-processing to achieve a visually cohesive presentation. The virtual characters were positioned 1.5 meters away from participants, within the social zone [74]. This placement aimed to support natural conversational perception [29] while ensuring gestures were clearly visible and easy to observe.

3.4.1 Data Selection. The animations for this study were sourced from the BEAT dataset [49], a large-scale motion capture repository known for its detailed audio, text, and emotion annotations. We reviewed actor performances in the dataset to ensure balanced stimuli, selecting based on: (i) gesture clarity, (ii) speech clarity, (iii) appropriate emotional expression, (iv) duration of approximately one minute, and (v) unaffected dialogue context. Each scenario we selected is a one-minute monologue, providing sufficient time for participants to form perceptions of the gestures, voice, and emotional expressions. To represent positive, neutral and negative emotions, we ensured the two gender-specific actors exhibited comparable intensity. Ultimately, three scenarios — one per emotion — featuring Wayne (male) and Ayana (female) were chosen for their alignment in style and expressiveness, meeting our criteria and enhancing study validity.

3.4.2 Speech Conditions. Two speech conditions were employed: natural and synthetic. Natural speech samples were obtained from the BEAT dataset, while synthetic speech was generated using ElevenLabs[3], a leading AI text-to-speech system chosen for its high-quality and realistic outputs [10]. ElevenLabs' extensive voice library enabled the selection of voices (female: Nala, male: Luke) that closely matched the original actors in accent and tone, utilizing default settings with text directly from the BEAT dataset.

Although we used only one TTS system, ElevenLabs was chosen for its perceptual realism and broad adoption. Prior work shows that user preferences for natural speech are consistent across TTS engines and immersion levels [15, 23, 38], making it a reliable proxy for comparing synthetic and natural voices.

Synthetic speech was generated using text transcripts processed through the TTS engine with embedded prompts for pauses. Temporal misalignments caused desynchronization with gestures, which were corrected in Adobe Audition by segmenting speech based on natural pauses and adjusting timing, pace and pause durations. This ensured synchronization with gestures and preserved the original performance's timing.

3.4.3 Gesture Conditions. We used two motion conditions: synthetic gestures generated by AMUSE [13] and natural gestures from the BEAT Dataset. AMUSE is an emotional speech-driven animation framework based on conditional latent diffusion [68], capable of synthesizing gestures from raw audio with explicit emotion control. We selected AMUSE due to its state-of-the-art emotion-aware performance and flexible editing features, which separate speech-driven content, emotion, and style into three latent vectors. Although only one model was employed, prior research indicates that perceptual differences largely stem from immersion context rather than the specific gesture generation method [20]. Hence, AMUSE adequately serves as a representative tool for examining the impact of natural versus synthetic gestures on user perception.

In our implementation, a vision transformer (ViT) [24] extracts audio features for the latent diffusion model, enabling content from one audio source (a_1) to be merged with the emotion and style from another (a_2). Specifically, we generated happy, sad and neutral gestures by combining Wayne and Ayana's speech with emotional profiles from Scott, Solomon and Miranda respectively. The resulting gestures were imported into Blender via the SMPL-X add-on [59], exported as BVH and retargeted in Auto-Rig Pro[4] to a CC4-based character in iClone. We then stitched 10-second AMUSE segments to form a continuous motion sequence. Finally, to keep consistency with AMUSE-generated gestures, we adapted the natural gestures from the BEAT dataset by removing their lower-body and finger motions to ensure consistency between conditions.

3.4.4 Lip Synchronization and Eye Blinks. Facial animations from the BEAT dataset were excluded because they were synchronized with recorded speech, leading to mismatches with our TTS-generated audio and reducing realism. Using a speech-to-face model could introduce artifacts and distract from our focus on gesture and speech. Consistent with prior gesture evaluation protocols [45], we opted for minimal facial animation, limited to lip synchronization and eye blinks, to avoid influencing participant judgments. Lip synchronization was achieved using AccuLIPS within iClone and enhanced with text transcripts to improve alignment accuracy. Eye-blinking was manually animated based on established conversational Spontaneous Eye-Blink Rates (SEBR) range (10.5–32.5 blinks/min) [25], ensuring realistic and consistent visual behaviour across stimuli.

3.4.5 Building VR and 2D Projects. Our study utilized two immersion conditions, VR and 2D, with separate Unity versions to ensure consistent design, rendering and experimental setups. In both

[2]https://www.reallusion.com/character-creator/.
[3]https://elevenlabs.io/
[4]https://lucky3d.fr/auto-rig-pro/doc/auto_rig.html

conditions, participants did not have a virtual presence and could not move within the environment: VR participants observed the character while standing in a fixed position, and 2D participants experienced a fixed camera view. Eye level for all participants was standardized and consistent across both conditions, as were rendering and performance settings. A 90 FPS frame limit was applied to both VR and 2D standalone exports, and lighting was pre-baked with no real-time lighting. Scenarios were randomized in both conditions to avoid order effects.

For the VR condition, OpenXR and Unity's XR Interaction Toolkit were employed. Since participants could naturally rotate or tilt their heads, A 'look-at' function was implemented, allowing the character to track the participant's head position. A VR questionnaire was embedded to maintain immersion. In the 2D condition, the camera's Field of View (FOV) and positioning were calibrated to replicate the VR perspective, ensuring consistent gesture visibility. Although screen sizes varied for prolific participants, the character remained centered with a constant FOV, ensuring consistent appearance and position while only affecting the periphery of the virtual environment. Here, the character faced the fixed camera, removing the need for a dynamic "look-at" function. An online questionnaire was embedded to collect responses (See supplementary video).

4 Results

The collected data was tested for normality using the Shapiro-Wilk test, which indicated deviations from a normal distribution. We therefore used the non-parametric Aligned Rank Transform (ART) [81] for analysis of a 3×4×2 mixed ANOVA with Gesture Voice Combinations (GVC) and Immersion as between-subject factors. Post-hoc analyses used Bonferroni-corrected ART-C contrasts [27].

Speech-Gesture Match: Our analysis revealed a significant main effect of GVC ($F(3, 211) = 57.43$, $p < .001$) for Speech-Gesture Match , as illustrated in Fig.2. Post hoc analysis showed that the GnVn condition produced higher match scores than all other conditions ($p < .001$. for all). Additionally, the GnVs condition outperformed both the GsVn ($p < .001$) and GsVs ($p < .001$) conditions, while no significant difference was observed between GsVn and GsVs. Immersion also had a significant effect ($F(1, 211) = 19.96$, $p < .001$), with VR resulting in perceived speech-gesture higher match scores than the 2D ($p < .001$). Furthermore, Scenario had a significant effect ($F(2, 422) = 5.18$, $p < .01$), with positive scenarios achieving better match scores than neutral scenarios ($p < .01$) (see Fig.3).

A significant interaction effect between GVC and Immersion was observed ($F(3, 211) = 7.47, p < 0.001$), as shown in Fig.4. The GnVn condition was perceived as significantly better matched in VR compared to 2D ($p < .001$). In both 2D and VR, GnVn differed significantly from GsVs and GsVn ($p < .001$ for all), with large effect size in VR ($r = .51$ for GsVn; $r = .61$ for GsVs). While no significant difference was found between GnVn and GnVs in 2D, GnVn scored significantly higher than GnVs in VR ($p < .001$). For synthetic combinations, VR slightly amplified perceived match for GnVs, with a significant distinction between GnVs and GsVs in both 2D and VR ($p < .001$ for all) and a increased effect size in VR ($r = .24$ for 2D, $r = .42$ for VR). In 2D, GnVs and GsVn did not differ significantly, but in VR, GnVs outperformed GsVn ($p < .01$).

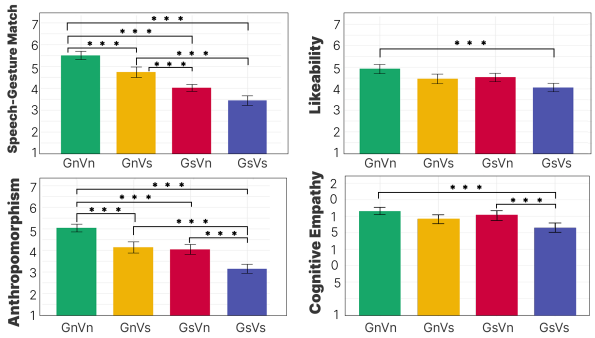


Figure 2: Main effect of GVC on various scales; fully natural combinations (GnVn) outperform synthetic combinations (GsVs) across all metrics.

Likeability: The analysis revealed three significant main effects. First, GVC significantly influenced likeability ($F(3, 211) = 6.82$, $p < .001$), with GnVn yielding higher likeability scores compared to GsVs ($p < .001$), as shown in Fig.2. Immersion also demonstrated a significant main effect ($F(1, 211) = 12.70$, $p < .001$), with participants in the VR condition reporting higher likeability than those in the 2D condition ($p < .001$). Furthermore, the Scenario factor was significant ($F(2, 422) = 24.62$, $p < .001$), with negative scenarios rated as less likeable than neutral scenario ($p < .001$) and positive scenario ($p < .001$), as illustrated in Fig.3.

Anthropomorphism: We saw three significant main effects. GVC significantly influenced Anthropomorphism ($F(3, 211) = 22.73$, $p < .001$), with GnVn producing higher Anthropomorphism scores than all synthetic combinations (All: $p < .001$) (see Fig.2). Additionally, GsVn and GnVs were rated significantly more human-like than GsVs (All: $p < .001$). Immersion demonstrated a significant main effect ($F(1, 211) = 17.00$, $p < .001$), with VR resulting in higher Anthropomorphism scores than the 2D immersion ($p < .001$). Furthermore, the scenario factor was significant ($F(2, 422) = 3.09$, $p < .05$), with positive scenarios receiving higher Anthropomorphism ratings than neutral scenarios ($p < .05$), as shown in Fig.3.

Importantly, we also found a significant interaction between GVC and Immersion ($F(3, 211) = 3.42$, $p < .05$) (see Fig.4). Post hoc contrasts indicated that participants exposed to VR rated the GnVn with higher anthropomorphism compared to the 2D condition ($p < .05$). Specifically, GnVn differed significantly from GsVs in both 2D ($p < .001$) and VR ($p < .001$), with an increased effect size observed in VR ($r = .28$ for 2D, $r = .46$ for VR). GnVn also received significantly higher human-likeness ratings than GsVn in VR ($p < .01$), something not observed in 2D. This suggests that VR's immersive qualities make realism-related differences more pronounced. Interestingly, the GnVs (natural gesture and synthetic speech) was also positively influenced by VR, showing higher anthropomorphism ratings compared to the 2D display ($p < .01$).

Co-Presence: Our analysis of Co-Presence data revealed a significant effect of immersion ($F(1, 211) = 43.48$, $p < .001$), consistent with our expectation that VR enhances the sense of co-presence ($p < .05$). Scenario significantly affected co-presence

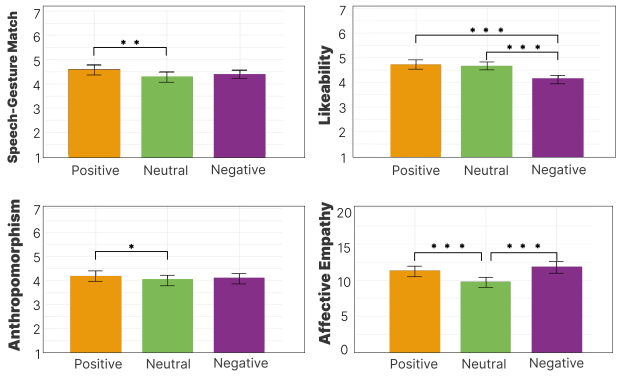


Figure 3: Main effect of Scenario on various scales: positive scenarios consistently received the highest ratings across all metrics.

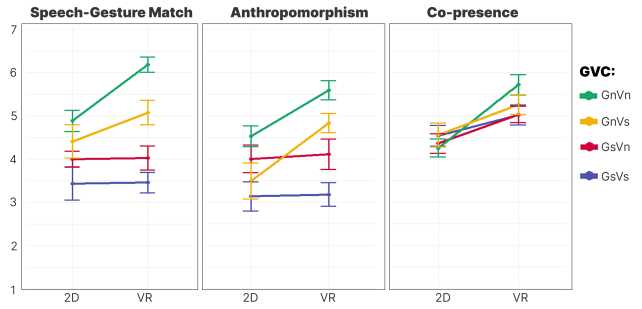


Figure 4: Interaction effect between GVC and Immersion on various scales; VR enhances natural combinations (GnVn) and accentuates differences with synthetic ones involving synthetic gestures.

($F(2, 422) = 4.15$, $p < .05$), with positive emotion scenarios leading to higher co-presence than neutral ($p < .05$). Additionally, The interaction between GVC and Immersion was significant, ($F(3, 211) = 2.98$,$p < .05$) (see Fig.4). Post-hoc comparisons revealed that VR significantly enhanced co-presence for GnVn ($p < .001$).

Emotional Responses: We did not find a significant main effect of GVC on either valence or arousal. A significant main effect of immersion on valence was found ($F(1, 211) = 18.10$, $p < .001$), with VR eliciting higher positive emotion than 2D ($p < .001$). Additionally, a significant interaction emerged between GVC and Scenario for both emotion valence ($F(6, 422) = 3.33$, $p < .001$) and emotional arousal ($F(6, 422) = 2.29$, $p < .05$), as displayed in Fig.5. Post hoc comparisons showed that in GnVn, GsVn, and GnVs, Positive scenarios elicited higher valence than Neutral and Negative scenarios ($p < .01$ for all). In GsVs, Positive scenarios had higher valence than Negative ($p < .001$) but did not differ from Neutral. Additionally, arousal in GsVs remained consistent across all scenarios.

Empathy Metrics: We found a main effect of GVC on Cognitive Empathy only ($F(3, 211) = 5.52$, $p < .01$) (see Fig.2). Pairwise comparisons indicated that synthetic gestures paired with natural voice (GnVn and GsVn) elicited higher empathy than the fully synthetic GsVs (GnVn:$p < .01$, GsVn: $p < .05$). Additionally, immersion had a significant main effect across all empathy dimensions (Cognitive: $F(1, 211) = 9.65$, $p < .01$; Affective: $F(1, 211) = 4.12$, $p < .05$; Compassionate: $F(1, 211) = 14.55$, $p < .001$), indicating that VR significantly enhances participants' empathy toward characters. The scenario exhibited a significant main effect across all dimensions of empathy ($p < .01$ for all), consistent with findings from previous research [37].

A significant interaction effect between immersion and scenario was observed for both cognitive empathy ($F(2, 422) = 8.81$, $p < .001$) and compassionate empathy ($F(2, 422) = 18.16$, $p < .001$), as visualized in Fig.6. In VR, both cognitive and compassionate empathy scores were significantly higher than in 2D for positive and neutral scenarios ($p < .05$ to $p < .001$), while no significant differences were observed in negative scenarios.

5 Discussion

Overall, our results support the central premise that synthetic modalities elicit dampened user responses compared to natural, motion-captured gestures and recorded human speech, and that VR intensifies these perceptions more than traditional 2D displays.

Speech–Gesture Match, Likeability, and Anthropomorphism: We observed similar patterns across all measures. Essentially, using both natural gestures and natural speech (GnVn) produced the highest ratings, while fully synthetic elements consistently lowered these perceptions. These results support H1 and corroborate findings by Ferstl et al. [31]. VR further boosted likeability across all gesture–voice combinations, contrasting with the work of Adkins et al. [1], who reported minimal VR benefits in the absence of audio and with masked facial expressions, and observed no significant effects on social presence. In our study, however, the inclusion of extra audio information and contextually rich cues (e.g., emotional expressions) possibly heightens the sense of physical presence and engagement in VR.

Importantly, immersion level significantly influenced perceived congruence and anthropomorphism. VR heightened the perceived compatibility and human-likeness of fully natural (GnVn) combinations, while magnifying the differences when synthetic gestures were used, aligning with H2 and prior work [52]. This suggests that greater spatial presence in VR may increase sensitivity to gesture quality, thus accentuating contrasts between natural-human and synthetic gestures. By contrast, the negative impact of synthetic voice diminished when paired with a natural-human gesture in VR. In these cases, sufficiently realistic visual cues appeared to compensate for the speech–gesture mismatch, consistent with the brain's tendency to prioritize visual information [73].

Our findings show that VR increases sensitivity to synthetic gestures, contrary to the study by Deichler et al. [20]. This difference may be attributed to their use of non-human avatars, limited emotional expression, and a small sample size. Although modern deep learning-based methods can generate highly realistic gestures [54], their evaluations predominantly occur in 2D, where performance is compared to human "ground truth" [45]. In extended reality (XR), congruence and plausibility remain critical [46], as

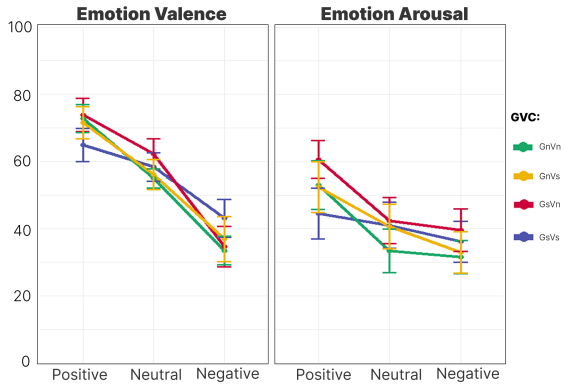


Figure 5: Interaction effect between GVC and Scenario on Emotion Engagement: the fully synthetic combination failed to differentiate positive and neutral valence and showed no arousal differences across scenarios.

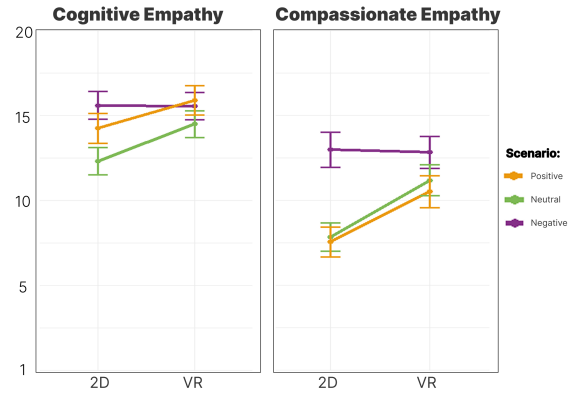


Figure 6: Interaction effect between Immersion and Scenario on Empathy: VR boosts empathy, except cognitive and compassionate empathy in negative scenarios.

mismatches in realism—especially among groups of virtual characters—can negatively affect user immersion [51]. Consequently, improving synthetic gesture quality in VR is paramount to achieving a more unified and immersive experience.

Lastly, positive scenarios garnered higher likeability scores than neutral or negative ones, reflecting a clear rank in user preference and corroborating findings from Higgins et al [37]. One plausible explanation for higher ratings of match and human likeness in positive scenarios is that the motion capture dataset—designed initially for gesture generation—included somewhat exaggerated gestures, which proved more effective at conveying positive affect.

Co-presence: In line with H2, immersion in VR enhanced users' sense of co-presence. In particular, VR enhances natural gesture-voice combinations more than fully synthetic ones. This aligns with previous research linking virtual human realism in animation, appearance, and voice to greater co-presence [12, 82, 87]. Contrary to prior findings that TTS voices in VR reduce presence due to mismatched realism and low appeal [38], our results show no significant differences in co-presence across gesture-voice combinations in 2D and VR. This divergence is attributed to our new finding that VR enhances the likeability of all gesture-voice combinations, supporting the established link between appeal and social presence [55]. Finally, the positive scenario demonstrated a notable increase in co-presence, likely attributable to a higher degree of speech-gesture congruence and enhanced anthropomorphic qualities. In summary, our findings indicate that natural gesture-voice combinations substantially enhance users' sense of co-presence, especially in immersive VR environments.

Emotional Responses: Firstly, the use of VR significantly enhanced the participants' experience of positive emotions, consistent with previous findings [22]. This effect can be attributed to the highly immersive nature of VR, which surpasses other forms of media [71]. We also found that gesture-voice combinations (GVCs) and scenarios jointly influenced emotional responses. Regarding emotional valence, only the fully synthetic combination failed to exhibit significant differentiation between positive and neutral scenarios. Similarly, the emotional arousal elicited by the fully synthetic combination did not vary significantly across scenario types. Contrary to H3, natural combinations did not evoke stronger emotional responses in positive scenarios compared to synthetic variants. Our new results suggest that synthetic combinations were perceived as more likeable and human-like in happy contexts, and these enhanced perceptions likely contributed to the stronger emotional responses observed. These results further demonstrated that only the fully synthetic pairing was less effective in evoking emotional responses, providing support for our H1. This aligns with previous findings that medium intensity gestures and moderate pitch variation were most effective in eliciting positive emotional responses in public speaking [67], suggesting that naturalistic nonverbal cues play a critical role in emotional engagement.

These results imply that scenario design and higher immersion are key to fostering emotional engagement with virtual characters. While mixing natural and artificial elements can be effective, relying entirely on artificial cues may reduce the emotional impact. This insight is particularly valuable for immersive VR applications, as narrative-driven interactive experiences have been shown to effectively evoke emotional responses [40]. Given that virtual characters often serve as primary conduits for narratives and social interactions [66], the integration of semi-synthetic or natural cues—rather than an exclusive reliance on fully synthetic designs—can significantly enhance their emotional impact and overall effectiveness.

Empathy: We found that GnVn significantly enhances cognitive empathy, which involves understanding others' thoughts and emotions [60], supporting H1. This effect may stem from the precise cues provided by natural gestures and voice, aligning with our new result that synthetic stimuli dampen emotional responses, potentially explaining the decrease in empathy. In contrast, affective and compassionate empathy, which rely more on emotional resonance, remain unaffected [18]. Notably, GnVn and GsVn outperform GsVs, with all effective combinations incorporating a natural voice. These results suggest that natural voice plays a critical role in maintaining cognitive empathy, which partially supports H1. The superior performance of GsVn in cognitive empathy tasks potentially arises

from the natural voice's richer verbal and emotional cues compensating for synthetic gestures, while GnVs appears more neutral due to prioritization of visual cues. In contrast to Higgins' findings [38], which suggested that TTS does not influence empathy, we attribute this discrepancy to the brevity of their scenario—only a few sentences long—which may have been insufficient to elicit meaningful empathetic responses regardless of voice type.

VR significantly enhances empathy across all dimensions, aligning with prior research suggesting that its heightened presence and engagement amplify empathic responses [69]. However, cognitive and compassionate empathy showed no increase in negative scenarios, likely due to a ceiling effect on uniformly high scores. Additionally, the influence of emotional scenarios on three-dimensional empathy aligns with findings that highlight the critical role of emotional states in shaping empathy [37]. VR is widely utilized by researchers as a medium to elicit empathy for various purposes, particularly by enabling embodiment to interact with virtual characters [26]. While character production often relies on resource-intensive motion capture, our results suggest that synthetic audio cues can be an effective alternative under resource constraints, provided scenarios are carefully designed to maintain engagement.

6 Limitations and future work

This work has several noteworthy limitations. First, variation in participant hardware in the 2D condition—Prolific users employed diverse computer setups—contrasted with the standardized use of Meta Quest 3 headsets in the VR group, potentially leading to inconsistent visual and auditory experiences. Second, while our scenarios were balanced for gesture and speech clarity, their limited variety may have constrained engagement outcomes. Third, the digital characters used here did not reach maximum realism; more sophisticated models (e.g., MetaHuman[5]) could prompt different viewer responses.

A limitation of this study is the use of a single gesture synthesis framework (AMUSE) and one TTS engine (ElevenLabs) to represent synthetic modalities. However, prior work suggests that user preferences and perceptual patterns are consistent across systems within the same modality. For example, Deichler et al.[20] found stable gesture evaluations across 2D and VR, and preferences for natural voices remain consistent across TTS engines[15, 23, 38]. Nonetheless, this approach may be perceived as restrictive, and future work should examine a broader range of systems to enhance generalizability.

Another limitation is the mismatch in emotional expression: while our gesture model supported emotion editing, the TTS engine lacked emotional tone control, potentially causing vocal–gesture incongruence. Using emotion-capable TTS could better align modalities. Additionally, the fixed one-minute monologue limited real-time interaction, possibly obscuring subtler social responses.

Future investigations should seek to enhance diversity, expanding the participant pool beyond a western-university based population sample. Employing intermediate interfaces — such as a 3D-viewer that allows free movement but not full immersion — may also clarify whether graded levels of immersion produce meaningful differences in perception and co-presence. More realistic and varied digital humans, potentially with fully rigged facial expressions or alternative appearance styles, could provide deeper insights into uncanny valley thresholds and anthropomorphism. Introducing real-time user–avatar interactions, incorporating eye-tracking to measure user engagement, and exploring proxemics could more closely mimic everyday social encounters and thus yield stronger external validity. In addition, it is possible that gesture had a greater impact than natural speech because users are familiar with synthetic voices. It would be interesting to explore how familiarity with synthetic cues influences perception. We also plan to investigate the impact of individual elements of conversational gesture (e.g., gesture type, timing, activation level) to further guide the development of co-speech gesture synthesis.

[5]https://metahuman.unrealengine.com

7 Conclusion

In summary, our findings highlight the substantial advantages of combining natural gestures and speech to enhance users' perceptions of realism, likeability, anthropomorphism, and co-presence — benefits that are magnified in virtual reality (VR). Examining a wide range of different measures simultaneously, Our study is first to identify that the gap in perceived quality between natural and synthetic gestures becomes particularly pronounced in VR, emphasizing the need to develop more lifelike synthetic gestures for immersive applications. Moreover, our results emphasize the key role of both scenario context and immersion in driving emotional engagement and empathy. While fully synthetic cues tend to diminish emotional resonance and reduce cognitive empathy, semi-synthetic combinations can still evoke strong affective responses — especially when enhanced by natural voices that convey richer verbal and emotional information.

We present these implications to highlight how tailoring vocal and gestural cues, alongside immersive platforms, can optimize both emotional engagement and empathy in virtual scenarios. Most gesture evaluation studies to date have used limited avatar representations, scenario contexts, and immersion environments to explore the relationship between gesture and voice. Our study suggests that gesture generation researchers should note the magnified gap between synthetic and natural gestures in VR and account for it in future designs and evaluations. Additionally, a single natural cue—either in voice or gesture—is sufficient to significantly boost immersion and empathy, offering a practical direction for developers working with resource-constrained systems. As VR and AI technologies rapidly evolve, further research into how best to coordinate voice and gesture is essential for creating immersive, engaging, and emotionally resonant virtual interactions.

Acknowledgments

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

[1] Alex Adkins, Aline Normoyle, Lorraine Lin, Yu Sun, Yuting Ye, Massimiliano Di Luca, and Sophie Joerg. 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.

[2] Maryam Alimardani, Robyn de Roode, Julija Vaitonyte, and Max M Louwerse. 2024. Effect of a Virtual Agent's Appearance and Voice on Uncanny Valley and

Trust in Human-Agent Collaboration. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*. 1–7.
[3] 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.
[4] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. 2022. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–19.
[5] Tenglong Ao, Zeyi Zhang, and Libin Liu. 2023. Gesturediffuclip: Gesture diffusion model with clip latents. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–18.
[6] Huda Barakat, Oytun Turk, and Cenk Demiroglu. 2024. Deep learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources. *EURASIP Journal on Audio, Speech, and Music Processing* 2024, 1 (2024), 11.
[7] Alberto Betella and Paul FMJ Verschure. 2016. The affective slider: A digital self-assessment scale for the measurement of human emotions. *PloS one* 11, 2 (2016), e0148037.
[8] Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. 2021. Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2027–2036.
[9] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. 2021. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE virtual reality and 3D user interfaces (VR)*. IEEE, 1–10.
[10] Martin Brehmer and Vito Buonassisi. 2024. Educators' Friend-Applying Generative AI to Create Effective Digital Learning Objects for Information Security Education: Toward Initial Design Principles. (2024).
[11] Julia Cambre, Jessica Colnago, Jim Maddock, Janice Tsai, and Jofish Kaye. 2020. Choice of voices: A large-scale evaluation of text-to-speech voice quality for long-form content. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[12] Emna Chérif and Jean-François Lemoine. 2019. Anthropomorphic virtual assistants and the reactions of Internet users: An experiment on the assistant's voice. *Recherche et Applications en Marketing (English Edition)* 34, 1 (2019), 28–47.
[13] 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.
[14] Kiran Chhatre, Renan Guarese, Andrii Matviienko, and Christopher Peters. 2025. Evaluating Speech and Video Models for Face-Body Congruence. In *Companion Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games (I3D Companion '25)*. Association for Computing Machinery, New York, NY, USA, Article 5, 3 pages.
[15] Erin K Chiou, Noah L Schroeder, and Scotty D Craig. 2020. How we trust, perceive, and learn from virtual humans: The influence of voice quality. *Computers & Education* 146 (2020), 103756.
[16] Mingyuan Chu and Peter Hagoort. 2014. Synchronization of speech and gesture: Evidence for interaction in action. *Journal of Experimental Psychology: General* 143, 4 (2014), 1726.
[17] Scotty D Craig and Noah L Schroeder. 2018. Design principles for virtual humans in educational technology environments. In *Deep comprehension*. Routledge, 128–139.
[18] Benjamin MP Cuff, Sarah J Brown, Laura Taylor, and Douglas J Howat. 2016. Empathy: A review of the concept. *Emotion review* 8, 2 (2016), 144–153.
[19] Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. 2023. Emotional Speech-Driven Animation with Content-Emotion Disentanglement. In *SIGGRAPH Asia 2023 Conference Papers* (Sydney, NSW, Australia) *(SA '23)*. Association for Computing Machinery, New York, NY, USA, Article 41, 13 pages.
[20] Anna Deichler, Jonas Beskow, and Axel Wiebe Werner. 2024. Gesture Evaluation in Virtual Reality. In *GENEA: Generation and Evaluation of Non-verbal Behaviour for Embodied Agents Workshop 2024*.
[21] Anna Deichler, Kiran Chhatre, Christopher Peters, and Jonas Beskow. 2021. Spatio-temporal priors in 3D human motion.
[22] Amir Dirin and Teemu H Laine. 2023. The influence of virtual character design on emotional engagement in immersive virtual reality: The case of feelings of being. *Electronics* 12, 10 (2023), 2321.
[23] Tiffany D Do, Ryan P McMahan, and Pamela J Wisniewski. 2022. A new uncanny valley? The effects of speech fidelity and human listener gender on social perceptions of a virtual-human speaker. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–11.
[24] Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
[25] 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.
[26] Haoyang Du, Songkai Jia, Joel Gautschi, Julia Quehenberger, David Lätsch, and Xueni Pan. 2023. Towards more child safety-oriented decisions through VR. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 687–688.
[27] Lisa A Elkin, Matthew Kay, James J Higgins, and Jacob O Wobbrock. 2021. An aligned rank transform procedure for multifactor contrast tests. In *The 34th annual ACM symposium on user interface software and technology*. 754–768.
[28] Cathy Ennis, Ludovic Hoyet, Arjan Egges, and Rachel McDonnell. 2013. Emotion capture: Emotionally expressive characters for games. In *Proceedings of motion on games*. 53–60.
[29] Cathy Ennis and Carol O'Sullivan. 2012. Perceptually plausible formations for virtual conversers. *Computer Animation and Virtual Worlds* 23, 3-4 (2012), 321–329.
[30] Elodie Etienne, Marion Ristorcelli, Sarah Saufnay, Aurélien Quilez, Rémy Casanova, Michaël Schyns, and Magalie Ochs. 2024. A Systematic Review on the Socio-affective Perception of IVAs' Multi-modal behaviour. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*. 1–10.
[31] 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.
[32] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F Troje, and Marc-André Carbonneau. 2023. ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech. In *Computer Graphics Forum*, Vol. 42. Wiley Online Library, 206–216.
[33] Ikhsanul Habibie, Mohamed Elgharib, Kripasindhu Sarkar, Ahsan Abdullah, Simbarashe Nyatsanga, Michael Neff, and Christian Theobalt. 2022. A motion matching-based framework for controllable gesture synthesis from speech. In *ACM SIGGRAPH 2022 conference proceedings*. 1–9.
[34] Chad Harms and Frank Biocca. 2004. Internal consistency and reliability of the networked minds measure of social presence. In *Seventh annual international workshop: Presence*, Vol. 2004. Universidad Politecnica de Valencia Valencia.
[35] Yuan He, André Pereira, and Taras Kucherenko. 2022. Evaluating data-driven co-speech gestures of embodied conversational agents through real-time interaction. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*. 1–8.
[36] Darragh Higgins, Donal Egan, Rebecca Fribourg, Benjamin Cowan, and Rachel McDonnell. 2021. Ascending from the valley: Can state-of-the-art photorealism avoid the uncanny?. In *ACM Symposium on Applied Perception 2021*. 1–5.
[37] Darragh Higgins, Yilin Zhan, Benjamin R Cowan, and Rachel McDonnell. 2023. Investigating the effect of visual realism on empathic responses to emotionally expressive virtual humans. In *ACM Symposium on Applied Perception 2023*. 1–7.
[38] 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.
[39] 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.
[40] Shafaq Irshad and Andrew Perkis. 2020. Increasing user engagement in virtual reality: The role of interactive digital narratives to trigger emotional responses. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. 1–4.
[41] Navdeep Kaur and Parminder Singh. 2023. Conventional and contemporary approaches used in text to speech synthesis: A review. *Artificial Intelligence Review* 56, 7 (2023), 5837–5880.
[42] Gaoussou Youssouf Kebe, Mehmet Deniz Birlikci, Auriane Boudin, Ryo Ishii, Jeffrey M Girard, and Louis-Philippe Morency. 2024. GeSTICS: A Multimodal Corpus for Studying Gesture Synthesis in Two-party Interactions with Contextualized Speech. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*. 1–10.
[43] Mark L Knapp, Judith A Hall, and Terrence G Horgan. 1978. *Nonverbal communication in human interaction*. Vol. 1. Holt, Rinehart and Winston New York.
[44] Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2020. The GENEA Challenge 2020: Benchmarking gesture-generation systems on common data. (2020).
[45] 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.
[46] Marc Erich Latoschik and Carolin Wienrich. 2022. Congruence and plausibility, not presence: Pivotal conditions for XR experiences and effects, a novel approach. *Frontiers in Virtual Reality* 3 (2022), 694433.
[47] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. 2021. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE/CVF International*

Conference on Computer Vision. 11293–11302.
[48] 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.
[49] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European conference on computer vision*. Springer, 612–630.
[50] 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.
[51] David Mal, Nina Döllinger, Erik Wolf, Stephan Wenninger, Mario Botsch, Carolin Wienrich, and Marc Erich Latoschik. 2024. Am I the odd one? Exploring (in) congruencies in the realism of avatars and virtual others in virtual reality. *Frontiers in Virtual Reality* 5 (2024), 1417066.
[52] David Mal, Erik Wolf, Nina Döllinger, Mario Botsch, Carolin Wienrich, and Marc Erich Latoschik. 2024. From 2D-Screens to VR: Exploring the Effect of Immersion on the Plausibility of Virtual Humans. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.
[53] Christos Mousas, Dimitris Anastasiou, and Ourania Spantidi. 2018. The effects of appearance and motion of virtual characters on emotional reactivity. *Computers in Human behavior* 86 (2018), 99–108.
[54] 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.
[55] Catherine S Oh, Jeremy N Bailenson, and Gregory F Welch. 2018. A systematic review of social presence: Definition, antecedents, and implications. *Frontiers in Robotics and AI* 5 (2018), 409295.
[56] Kunkun Pang, Dafei Qin, Yingruo Fan, Julian Habekost, Takaaki Shiratori, Junichi Yamagishi, and Taku Komura. 2023. Bodyformer: Semantics-guided 3d body gesture synthesis with transformer. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–12.
[57] Dhaval Parmar, Stefan Olafsson, Dina Utami, Prasanth Murali, and Timothy Bickmore. 2022. Designing empathic virtual agents: manipulating animation, voice, rendering, and empathy to create persuasive agents. *Autonomous agents and multi-agent systems* 36, 1 (2022), 17.
[58] Prabesh Paudel, Hyung-Jeong Yang, Anjitha Divakaran, Seung-won Kim, JI-eun Shin, and Soo Hyung Kim. 2024. A Survey on Identifying Empathic Expression in Remote Collaboration from Empathic Computing. *IEEE Access* (2024).
[59] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.
[60] Philip A Powell and Jennifer Roberts. 2017. Situational determinants of cognitive, affective, and compassionate empathy in naturalistic digital interactions. *Computers in Human Behavior* 68 (2017), 137–148.
[61] Susanne Putze, Dmitry Alexandrovsky, Felix Putze, Sebastian Höffner, Jan David Smeddinck, and Rainer Malaka. 2020. Breaking the experience: Effects of questionnaires in vr user studies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.
[62] Xingqun Qi, Chen Liu, Lincheng Li, Jie Hou, Haoran Xin, and Xin Yu. 2024. Emotiongesture: Audio-driven diverse emotional co-speech 3d gesture generation. *IEEE Transactions on Multimedia* (2024).
[63] Chao Qu, Willem-Paul Brinkman, Yun Ling, Pascal Wiggers, and Ingrid Heynderickx. 2014. Conversations with a virtual human: Synthetic emotions and human responses. *Computers in Human Behavior* 34 (2014), 58–68.
[64] Mira Ram, Lorie Loeb, Michael Cohen, and James Mahoney. 2024. Virtual Avatar Movement Veracity and Its Effect on Social Interaction in Virtual Reality. In *Science and Information Conference*. Springer, 289–307.
[65] Imogen C Rehm, Emily Foenander, Klaire Wallace, Jo-Anne M Abbott, Michael Kyrios, and Neil Thomas. 2016. What role can avatars play in e-mental health interventions? Exploring new models of client–therapist interaction. *Frontiers in Psychiatry* 7 (2016), 186.
[66] Mark O Riedl and R Michael Young. 2003. Character-focused narrative generation for execution in virtual worlds. In *International Conference on Virtual Storytelling*. Springer, 47–56.
[67] Emma Rodero. 2022. Effectiveness, attractiveness, and emotional response to voice pitch and hand gestures in public speaking. *Frontiers in communication* 7 (2022), 869084.
[68] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
[69] Jon Rueda and Francisco Lara. 2020. Virtual reality and empathy enhancement: Ethical aspects. *Frontiers in Robotics and AI* 7 (2020), 506984.
[70] Ralph Schroeder. 2002. Social interaction in virtual environments: Key issues, common themes, and a framework for research. In *The social life of avatars: Presence and interaction in shared virtual environments*. Springer, 1–18.
[71] Mel Slater. 2009. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1535 (2009), 3549–3557.
[72] Rukshani Somarathna, Tomasz Bednarz, and Gelareh Mohammadi. 2022. Virtual reality for emotion elicitation–a review. *IEEE Transactions on Affective Computing* (2022).
[73] Dustin Stokes and Stephen Biggs. 2014. The dominance of the visual. *Perception and its Modalities* (2014), 350–378.
[74] Eric Sundstrom and Irwin Altman. 1976. Interpersonal relationships and personal space: Research review and theoretical model. *Human Ecology* 4 (1976), 47–67.
[75] Paul Taylor. 2009. *Text-to-speech synthesis*. Cambridge university press.
[76] Angela Tinwell, Mark Grimshaw, Debbie Abdel Nabi, and Andrew Williams. 2011. Facial expression of emotion and perception of the Uncanny Valley in virtual characters. *Computers in Human behavior* 27, 2 (2011), 741–749.
[77] Andreas Triantafyllopoulos, Björn W Schuller, Gökçe İymen, Metin Sezgin, Xiangheng He, Zijiang Yang, Panagiotis Tzirakis, Shuo Liu, Silvan Mertes, Elisabeth André, et al. 2023. An overview of affective speech synthesis and conversion in the deep learning era. *Proc. IEEE* 111, 10 (2023), 1355–1381.
[78] Matias Volonte, Yu-Chun Hsu, Kuan-Yu Liu, Joe P Mazer, Sai-Keung Wong, and Sabarish V Babu. 2020. Effects of interacting with a crowd of emotional virtual humans on users' affective and non-verbal behaviors. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 293–302.
[79] Hendric Voß and Stefan Kopp. 2023. Augmented Co-Speech Gesture Generation: Including Form and Meaning Features to Guide Learning-Based Gesture Synthesis. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*. 1–8.
[80] Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview. 209–232 pages.
[81] 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.
[82] Yanxiang Wu, Sabarish V Babu, Rowan Armstrong, Jeffrey W Bertrand, Jun Luo, Tania Roy, Shaundra B Daily, Lauren Cairco Dukes, Larry F Hodges, and Tracy Fasolino. 2014. Effects of virtual human animation on emotion contagion in simulated inter-personal experiences. *IEEE transactions on visualization and computer graphics* 20, 4 (2014), 626–635.
[83] Cindy Xu, Mengyu Chen, Pranav Deshpande, Elvir Azanli, Runqing Yang, and Joseph Ligman. 2024. Enabling Data-Driven and Empathetic Interactions: A Context-Aware 3D Virtual Agent in Mixed Reality for Enhanced Financial Customer Experience. In *2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 186–192.
[84] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2022. The GENEA Challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 736–747.
[85] Jie Zeng, Yoshiki Takahashi, Yukiko I Nakano, Tatsuya Sakato, and Hannes Högni Vilhjálmsson. 2024. Modifying Gesture Style with Impression Words. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*. 1–9.
[86] 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.
[87] Katja Zibrek, Elena Kokkinara, and Rachel McDonnell. 2017. Don't stand so close to me: investigating the effect of control on the appeal of virtual humans using immersion and a proximity-based behavioral task. In *Proceedings of the ACM symposium on applied perception*. 1–11.