Evaluation of generative models for emotional 3D animation generation in VR

Kiran Chhatre ⓘ *, Renan Guarese ⓘ, Andrii Matviienko ⓘ and Christopher Peters ⓘ

School of Electronic Engineering and Computer Science (EECS), KTH Royal Institute of Technology, Stockholm, Sweden

Introduction: Social interactions incorporate various nonverbal signals to convey emotions alongside speech, including facial expressions and body gestures. Generative models have demonstrated promising results in creating full-body nonverbal animations synchronized with speech; however, evaluations using statistical metrics in 2D settings fail to fully capture user-perceived emotions, limiting our understanding of the effectiveness of these models.

Methods: To address this, we evaluate emotional 3D animation generative models within an immersive Virtual Reality (VR) environment, emphasizing user—centric metrics-emotional arousal realism, naturalness, enjoyment, diversity, and interaction quality—in a real-time human-agent interaction scenario. Through a user study ($N = 48$), we systematically examine perceived emotional quality for three state-of-the-art speech-driven 3D animation methods across two specific emotions: happiness (high arousal) and neutral (mid arousal). Additionally, we compare these generative models against real human expressions obtained via a reconstruction-based method to assess both their strengths and limitations and how closely they replicate real human facial and body expressions.

Results: Our results demonstrate that methods explicitly modeling emotions lead to higher recognition accuracy compared to those focusing solely on speech-driven synchrony. Users rated the realism and naturalness of happy animations significantly higher than those of neutral animations, highlighting the limitations of current generative models in handling subtle emotional states.

Discussion: Generative models underperformed compared to reconstruction-based methods in facial expression quality, and all methods received relatively low ratings for animation enjoyment and interaction quality, emphasizing the importance of incorporating user-centric evaluations into generative model development. Finally, participants positively recognized animation diversity across all generative models.

1 Introduction

Conversational interactions between users and virtual characters are crucial for immersive social experiences in VR, requiring the generation of behaviors such as speech (Vasiliu et al., 2025), gestures (Ghorbani et al., 2023), and facial expressions (Kruzic et al., 2020). However, accurately replicating verbal and non-verbal cues remains challenging. Social interactions incorporate multiple non-verbal modalities–such as gesture arousal, facial expressions, eye contact, and body posture–that are vital for conveying emotions, often taking precedence over verbal language (De Stefani and De Marco, 2019; Sharkov et al., 2022). Moreover, non-verbal expressions guide human behavior by providing

key signals on how to respond to others (Stewart et al., 2024) and shape perceptions of personality traits (Tracy et al., 2015). Yet, misconceptions persist about how these non-verbal cues function in real conversations, making it difficult to confirm whether the generated animation in virtual characters truly conveys the intended emotional behavior. These challenges highlight the need for comprehensive deep learning models that account for the interplay among multiple modalities (Patterson et al., 2023).

In virtual environments, verbal and non-verbal expressions are essential for delivering immersive social experiences, contributing significantly to users' social presence and emotional engagement (Smith and Neff, 2018). The intricate interplay of verbal and non-verbal cues complicates both the modeling and evaluation of character behavior. Early research employed hand-crafted animation and rule-based models (Cassell et al., 2001; Poggi et al., 2005), but these methods cannot capture the full range of possible cues, limiting the fidelity of social interactions. More recent work has leveraged motion capture to create high-fidelity behavior for teleoperated (or Wizard of Oz; WOZ) avatars (Fraser et al., 2022; Zhang et al., 2023), which excel at conveying emotional, full-body expressions by relying on human performers. Yet, this approach is costly and less scalable due to expensive motion capture technology. With the rapid development and widespread use in generating speech and motion content, generative models offer new possibilities for creating human-like social agents. Tools such as text-to-speech (TTS) systems (Kim et al., 2021; Casanova et al., 2021; Vasiliu et al., 2025) and speech-to-animation models (Yi et al., 2023) have opened new avenues for building the virtual characters. By leveraging these models, one can automate their creation: TTS produces natural-sounding speech for dialogue scripts, while speech-to-animation models synchronize gestures and facial expressions with spoken words, adding emotional depth to interactions (Chhatre et al., 2024; Daněček et al., 2023a). Although recent studies demonstrate high performance in animation realism, expressiveness, and diversity in *monologue* scenarios (Fan et al., 2022; Chhatre et al., 2024; Daněček et al., 2023a), the effectiveness of these models in VR *dialogue* settings–where human interactions with virtual characters come into play–remains uncertain.

Existing generative models can produce full-body animations from speech (Chhatre et al., 2024; Daněček et al., 2023b; Ginosar et al., 2019; Alexanderson et al., 2023; Yang et al., 2023a,b) and provide holistic co-speech, full-body datasets (Mughal et al., 2024; Liu et al., 2024). However, gesture generation has largely relied on *objective metrics*—for instance, Frechet Gesture Distance (Maiorca et al., 2022; Yoon et al., 2020) (comparing latent features between generated and ground-truth motion), beat alignment (Li et al., 2021; Valle-Pérez et al., 2021) (assessing motion-speech correlation via kinematic and audio beats), semantic-relevance gesture recall (Liu et al., 2022b), and gesture diversity (Li et al., 2023; Liu et al., 2022a) (covering beat, deictic, iconic, and metaphoric gestures). While these metrics are useful, they often fail to capture how humans truly perceive gestures. *User-centered metrics*—including perceived emotional realism, naturalness, and diversity–remain underexplored, even though they are crucial for evaluating virtual characters in a social context (Chhatre et al., 2025).

The effectiveness of these models depends on how well users perceive expressed emotions and interactional effects during social exchanges.

Although some studies have evaluated virtual faces and gesture generation—for example, investigating the uncanny valley effect (Di Natale et al., 2023), the GENEA Challenge (Kucherenko et al., 2023) on speech-driven gesture generation in monadic and dyadic contexts, research on the relationship between empathy and facial-based emotion simulation in VR (Della Greca et al., 2024), and AV-Flow (Chatziagapi et al., 2025) for dyadic speech and talking-head generation–these typically focus on either gestures or virtual faces, rather than a holistic 3D perceptual experience combining both face and body. Chhatre et al. (2025) examine how integrating facial expressions with body gestures influences animation congruency and synchronize the generated motion with the driving speech; however, they do not address diverse or emotionally rich conversational contexts. Closer to our work, Deichler et al. (2024) evaluated animations generated by such models from a third-person viewpoint in monologue or dialogue, emphasizing the impact of an immersive VR environment relative to a 2D setting. However, their study did not address real-time human-virtual character interaction or emotional conversation contexts, leaving the effects of integrating face and body unclear. Consequently, subjective qualities remain insufficiently studied for a human-virtual character in dyadic emotional interaction.

In this work, we address this gap by evaluating generative models for animation in VR, focusing on immersive human-virtual human interactions within emotionally contextual dialogues. Our user study employs two arousal conditions–Happy (high arousal) and Neutral (mid arousal)–based on the circumplex model of affect (Russell, 1980). Focusing on these two states allows us to examine five key perceptual factors–realism, naturalness, enjoyment, diversity, and interaction quality (see Section 4)— without introducing excessive complexity into the study design. We additionally compare participant ratings with outputs from a pretrained deep-learning classifier trained on Ekmans eight basic emotions (Ekman, 1993), collapsing its predictions into the same two categories (happy vs. neutral) for consistency. Limiting the scope to happy and neutral makes the experiment tractable while establishing a foundation for future studies that may explore a broader range of affective states, including complex emotions such as guilt and embarrassment, as well as negative emotions such as anger and disgust. Our specific goal is to assess the perceptual impact of these two emotional conditions and, based on the findings, iteratively improve the study to better capture the perceived effects of more diverse affective states. Building on previous work (Heesen et al., 2024; Liu, 2024; Tan and Nareyek, 2009; Conley et al., 2018; Zhu et al., 2023), we concentrate on critical perceived emotional animation attributes. We assess these qualities through a VR-based user study ($N = 48$), designed to explore immersion and social presence during interaction with virtual characters, in contrast to 2D videos (Mal et al., 2024). Advances in interactive media now support qualitatively richer experiences; immersive VR, in particular, can influence users' physiology, psychology, behavior, and social responses (Lombard et al.,

2009). To investigate how high-quality, computer-generated speech-driven animations of virtual characters affect factors such as enjoyment, persuasion, and social relationship, we therefore conduct our study in a VR setting. The same methodology could later be adapted to mixed or augmented reality environments. This study highlights the importance of perceptual evaluation, as objective metrics alone cannot fully capture the validity of generated gestures. Moreover, integrating multiple generative models for real-time interaction–combining speech and 3D animation–offers a promising direction for computational interaction systems. Rather than exclusively training or refining new models, our approach emphasizes a holistic perceptual assessment of current models to guide future model development.

While numerous speech-driven face-expression and body-gesture generative models exist, we specifically chose three representative methods based on their state-of-the-art performance on objective metrics such as realism, diversity, Frechet Gesture Distance, and beat alignment, as reported by their original authors. In our implementation, we use the SMPL-X (Pavlakos et al., 2019) parametric model for representing virtual humans in 3D. The three chosen models—EMAGE (Liu et al., 2024), TalkSHOW (Yi et al., 2023), and AMUSE (Chhatre et al., 2024)–exhibited top performance, with AMUSE uniquely focusing on emotional 3D body gestures. To incorporate face animation with AMUSE, we employed FaceFormer (Fan et al., 2021), a SMPL-X-compatible, speech-driven face-expression model. Additionally, we compared these generative models to real human expressions by capturing a human performers 3D face and body via the PIXIE (Feng et al., 2021a) frame-level reconstruction method. The core of our study is a user-based evaluation that reveals the strengths and weaknesses of these generative approaches, as well as how they shape user perception in VR. By also comparing the generative outputs to reconstruction-based real expressions, we shed light on how closely current generative methods can replicate real human body and facial expressions. This evaluation informs the selection of models best suited for specific applications, depending on which attributes–such as realism, naturalness, enjoyment, diversity, or interaction quality–are most critical. Our main contributions are as follows:

- To the best of our knowledge, we present the first perceptual evaluation of generative models for emotional 3D animation in real-time human-virtual character interactions within an immersive VR environment.
- We conduct a VR-based user study ($N = 48$), evaluating three representative generative methods with demonstrated capabilities in emotional animation generation.
- We evaluate the realism, naturalness, enjoyment, diversity, and interaction quality of the generated animations and investigate their impact on user perception.

In Section 2, we review related work, and in Section 3, we cover key concepts and provide an overview of our implementation details. Section 4 details our user study, while Section 5 presents the results. Section6 discusses the findings and practical implications, followed by Section 7, which addresses limitations and future work. Finally, Section 8 concludes the paper.

2 Related work

2.1 Social interaction

Social interaction is a complex interplay of language, gestures, and other nonverbal behaviors. The theory of embodied cognition suggests that spoken language evolved from motor actions, with empirical studies showing motor system involvement in both language and gesture production and comprehension (Gentilucci et al., 2006; Rizzolatti and Arbib, 1998). Research indicates that gestures and spoken language function in sync during face-to-face communication, with symbolic gestures sometimes replacing verbal components (Andric et al., 2013). This synchronization reflects the interaction between the sensory-motor and language processing systems (Bernardis and Gentilucci, 2006; McNeill, 1992). Nonverbal behaviors–facial expressions, gestures, posture, and gaze–are essential for conveying intentions, often enhancing or replacing verbal communication to produce a more accurate display of emotions than any single channel alone (Gunter and Bach, 2004; Zhao et al., 2018). Gestures, in particular, are tightly integrated with speech (Özyürek, 2014; He et al., 2018). However, the intricate ways these modalities interact remain not fully understood, even in human studies, making it challenging to develop virtual agents that accurately replicate such interactions.

2.1.1 Emotions in social interaction

Emotion has been a central theme in social-interaction research for decades. De Stefani and De Marco (2019) argue that the human Mirror Mechanism associates language in shared sensorimotor representations, tightly coupling gestures, speech and affect; Huang and Lajoie (2023) show that co-regulation of such social-emotional exchanges is critical for effective collaborative learning; and Marinetti et al. (2011) treat emotions as dynamic, context-dependent processes, comparing the competence of humans with that of emotionally aware artificial agents.

Emotions encountered in these interactions can be cast either as discrete categories (Figure 1-left) or as points in a continuous affective space (Figure 1-right). Ekmans taxonomy lists six basic classes–anger, disgust, fear, happiness, sadness and surprise–assumed to be biologically hard-wired (Ekman, 1993). Dimensional models assign emotions in low-dimensional manifolds: Schlosberg (1954) organized facial expressions along pleasant-unpleasant and attention-rejection axes with activation as a third dimension, while the widely used circumplex model maps emotions onto arousal and valence axes whose origin denotes neutrality (Russell, 1980). Later work in the Vector model confirms the emotions are structured in terms of arousal and valence such that a positive valence represents appetitive motivation and negative valence represents defensive motivation (Bradley et al., 1992). The stability of positive versus negative affect in two separate systems is analyzed into the positive activation - negative activation model (Watson and Tellegen, 1985). Finally, Plutchik (2001) integrates categorical and dimensional views within a 3D framework where it arranges emotions in concentric circles, where inner circles are more basic and the outer circles are also formed by blending the inner circle emotions. Our study adopts the circumplex model, distinguishing mid- (neutral)
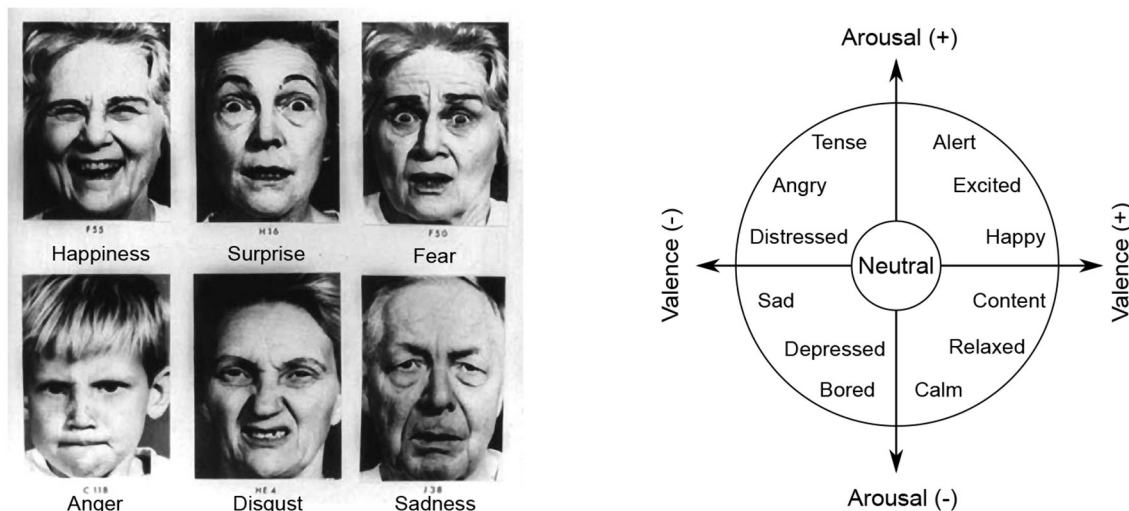
FIGURE 1
Emotion classification. **Left**: Ekmans discrete-emotion theory identifies six basic categories−anger, disgust, fear, happiness, sadness, and surprise−treating each as a distinct class rather than points on a continuum (Ekman, 1993). **Right**: The circumplex model (Russell, 1980) places emotions in a two-dimensional space spanned by arousal and valence; the center represents neutral arousal and neutral valence.

and high-arousal (happy) conditions, and augments participant judgements with an automatic emotion recognition deep learning model trained on an extended Ekman-style eight-class taxonomy, including contempt and neutral as additional categories.

2.2 VR-based interaction

In VR, interactions with virtual characters must be highly realistic to feel lifelike, a requirement with broad applications in entertainment and psychological research (Zhang et al., 2023). Approaches to creating these interactions typically fall into two categories. First, *rule-based models* rely on predefined rules or human interaction knowledge (Kopp et al., 2006; Cassell et al., 2001; Poggi et al., 2005), often using pre-recorded animations triggered by algorithms or manual intervention (Thiébaux et al., 2008; Marsella et al., 2013; Pan and Hamilton, 2018). These methods are constrained by limited motion variety, leading to repetitive behaviors (Zhang et al., 2023). Second, *teleoperation* (WOZ avatar approach) assigns human actors to drive virtual characters' voice and body movements (Fraser et al., 2022; Brandstätter and Steed, 2023; Zhang et al., 2023). Though highly realistic, this approach depends on expensive motion capture devices and restricts the number of actors who can simultaneously participate in a single VR experience. Some studies have explored using one human to control multiple virtual characters (Osimo et al., 2015; Yin et al., 2022; Brandstätter and Steed, 2023; Yin et al., 2024), but this reduces the variety of generated behaviors (Yin et al., 2022), limiting scalability for group interactions in VR.

Recently, industry applications have emerged for generating narrated avatar videos (Hedra, 2025; Synthesia, 2025; Microsoft Mesh, 2025; Soul Machines, 2025), 3D interactive non-player characters (Inworld, 2025; Convai, 2025; NVIDIA ACE, 2025), and user-interactable virtual characters (Replika, 2025). However, many

of these platforms lack flexibility and seamless integration with tools like Blender or Unity (Ton Roosendaal, 2025; Tim Sweeney, 2025; Unity Technologies, 2025), hindering direct comparison with rule-based or teleoperated methods.

Despite these challenges, numerous studies have examined conversational virtual characters in VR (Smith and Neff, 2018; Thomas et al., 2022; Herrera et al., 2018), focusing on aspects like rendering realism (Kokkinara and Mcdonnell, 2015; Zibrek et al., 2018; Patotskaya et al., 2023), animation realism (Guadagno et al., 2007; Rosenthal-von der Pütten et al., 2010), facial expressions and eye gaze (Roth et al., 2018a,b), body gestures (Huesser et al., 2021), subtle social cues (Reeves and Nass, 1996), and emotion disclosure (Barreda-Ángeles and Hartmann, 2021; Hancock et al., 2007). Yet most rely on rule-based or teleoperated animations, limiting both variety and quality of generated behaviors.

2.3 Generative models for virtual character interaction

Generative probabilistic models are widely used to produce speech and human motion. Recent advances in conditional constraints enable virtual social interactions with specific styles or emotions, offering low-cost, automated generation and diverse behaviors due to their probabilistic nature (Ma et al., 2025).

Recent methods employ deep neural networks to create motion animations, emphasizing convincing non-verbal behaviors. They generate 3D talking heads from speech (Pham et al., 2017a,b; Karras et al., 2017; Taylor et al., 2017; Zhou et al., 2018; Cudeiro et al., 2019; Richard et al., 2021; Fan et al., 2022; Xing et al., 2023) and synthesize 3D body gestures (Ginosar et al., 2019; Qi et al., 2023; Yoon et al., 2020; Habibie et al., 2022; Yang et al., 2023b). Some jointly produce body and facial animations via SMPL-X (Pavlakos et al., 2019; Yi et al., 2023), enabling

TABLE 1 Comparison of methods for 3D animation generation.

Method	Generative model type	Audio model type	Text
EMAGE	VQ-VAE	Content, rhythm-aware TCN	✓
TalkSHOW	Transformer (face), VQ-VAE (body)	Wav2Vec	✗
AMUSE+FaceFormer	Transformer (face), diffusion (body)	Content, emotion, style-aware ViT	✗

This table compares three generative models for gesture animation in VR. The models differ in their primary generative architecture for facial expressions and body gestures, as well as their approach to processing audio input. Additionally, one model accepts textual transcripts as supplementary input. All methods output body pose (θ) and facial expression (ψ) parameters.

more expressive behaviors. While speech-driven animation control remains underexplored, recent studies introduce motion style control (Yin et al., 2023; Alexanderson et al., 2023) and include style and emotion constraints (Fan et al., 2022; Chhatre et al., 2024).

For speech generation, text-to-speech (TTS) systems allow emotional variation in tone, pitch, and rhythm (Kim et al., 2021; Casanova et al., 2021), thereby enhancing user engagement in virtual interactions. Although individual models for speech and animation show promise, they are often developed and evaluated in isolation. In contrast, our approach integrates TTS and generative animation into a unified VR system, enabling a more comprehensive evaluation. We specifically examine how effectively they convey user perception of 3D full-body emotional responses and how these factors impact interaction quality in immersive environments.

3 Implementation details

3.1 Preliminaries: geometry, appearance, and rendering

We adopt the SMPL-X model (Pavlakos et al., 2019) to represent 3D body geometry, defined by $M(\beta, \theta, \psi)$. This model generates a mesh M from the identity shape $\beta \in \mathbb{R}^{300}$, pose $\theta \in \mathbb{R}^{J \times 3}$, and facial expression $\psi \in \mathbb{R}^{100}$, where J represents the number of body joints. For its appearance, we use SMPL-X UV coordinates, and the shaded textures are obtained by sampling albedo α, surface normals, and lighting. The Embodied Conversational Agent (Cassell, 2000) SMPL-X meshes–referred to as the "agent" hereafter–are animated in Blender using outputs from the generative models summarized in Table 1 and detailed in Section 3.2.

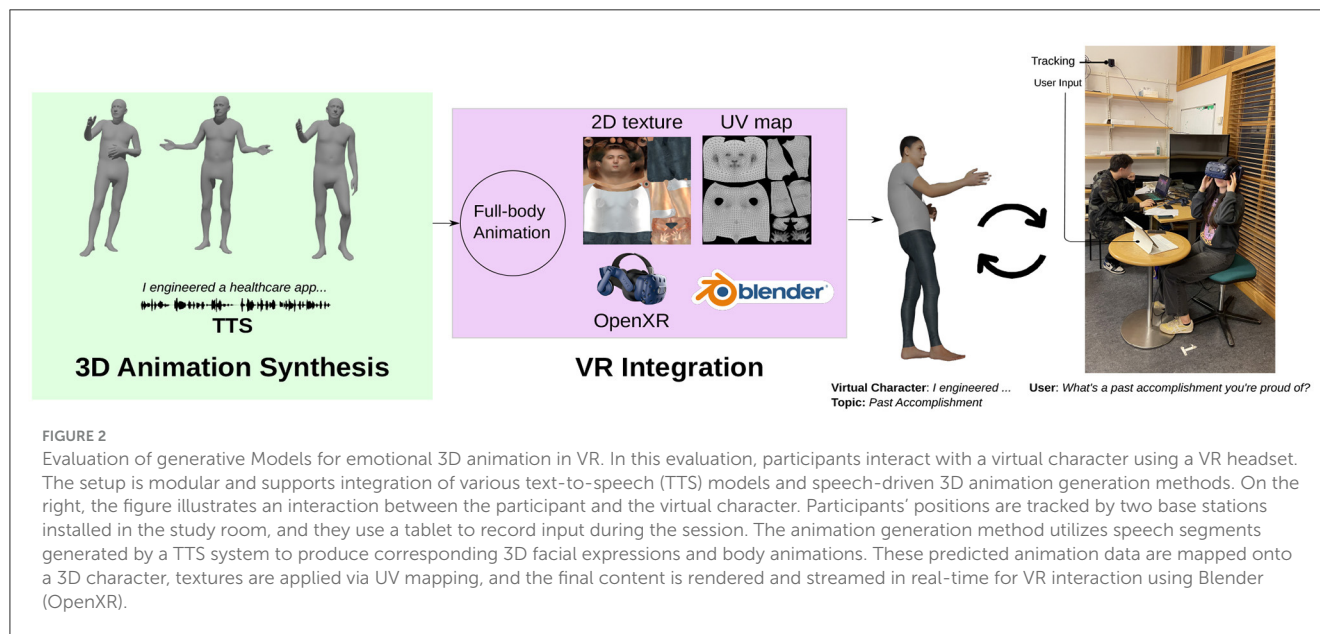### 3.2 Generative models

As shown in Figure 2, we fully synthesize a virtual characters motion and speech. We select three state-of-the-art models based on their performance in generating synthetic animations driven solely by audio input. These audio-based models generate 3D motion from speech and transcripts, and each has demonstrated strong speech-driven animation capabilities. In our pipeline, the driving speech–or video for the reconstruction baseline—is first fed to the selected model to predict full-body animation parameters. The resulting motion is then retargeted to a textured SMPL-X agent and placed in an outdoor Blender scene with appropriate lighting and camera placement. Finally, the animated scene is streamed to participants in real-time conversation through an HTC Vive

Pro 2 headset. We conduct quantitative evaluations comparing all models. Each method is applied to predefined scenarios with unique topics; transcripts and speech are generated via TTS, which then drive the 3D motion. The system is modular, allowing any component to be replaced as needed.

We utilize three state-of-the-art audio-driven generation models compatible with the SMPL-X mesh: EMAGE (Liu et al., 2024), TalkSHOW (Yi et al., 2023), and a combination of AMUSE (Chhatre et al., 2024) (for body) and FaceFormer (Fan et al., 2022) (for face). In Table 1, we summarize the specifics of each model. All models take raw audio as input and produce 3D animations. EMAGE and TalkSHOW output both ψ and θ parameters, whereas AMUSE outputs θ parameters and FaceFormer outputs ψ parameters; both parameter sets are integrated at the frame level after inference. Specifically, FaceFormer outputs meshes with the FLAME topology (Li et al., 2017). We convert these meshes into FLAME expression parameters by fitting the registered 3D mesh to the FLAME model using the FLAME fitting framework (Bolkart, 2013) and the Broyden-Fletcher-Goldfarb-Shanno optimizer. Once we obtain the ψ parameters, we combine them with the θ parameters– aligning jaw rotations framewise–to create a single motion file. Throughout this process, the identity parameters (β) from the original AMUSE output are preserved. Next, EMAGE accepts text transcripts as an additional input. All geometric parameters are passed to the SMPL-X Blender add-on, which imports the meshes into the Blender scene. Each imported SMPL-X mesh includes a shape-specific rig and blend shapes for shape, expression, and pose parameters. We use consistent sampled β parameters and an α texture across all models. All evaluated models–EMAGE (Liu et al., 2024), TalkSHOW (Yi et al., 2023), AMUSE (Chhatre et al., 2024), and FaceFormer (Fan et al., 2021)–were made publicly available by their respective authors. An introduction to each method is provided in the Supplementary Section 2.

The models process audio features differently. TalkSHOW uses a pre-trained Wav2Vec (Baevski et al., 2020) model to extract speech features, while EMAGE and AMUSE employ specialized models for this purpose. EMAGE uses a content- and rhythm-aware Temporal Convolutional Network (TCN) (Lea et al., 2017) that distinguishes gestures related to semantic content versus rhythm for each frame. FaceFormer also uses Wav2Vec to extract speech features, whereas AMUSE uses a Vision Transformer (ViT)-based model (Dosovitskiy, 2020). The AMUSE model additionally disentangles content-, emotion-, and style-aware features from the driving speech, explicitly modeling the impact of emotions on generated gestures. The backbone architectures used for gesture and expression generation vary among the models. EMAGE utilizes multiple Vector Quantized Variational AutoEncoders (VQ-VAE) (Van Den Oord et al., 2017)

FIGURE 2
Evaluation of generative Models for emotional 3D animation in VR. In this evaluation, participants interact with a virtual character using a VR headset. The setup is modular and supports integration of various text-to-speech (TTS) models and speech-driven 3D animation generation methods. On the right, the figure illustrates an interaction between the participant and the virtual character. Participants' positions are tracked by two base stations installed in the study room, and they use a tablet to record input during the session. The animation generation method utilizes speech segments generated by a TTS system to produce corresponding 3D facial expressions and body animations. These predicted animation data are mapped onto a 3D character, textures are applied via UV mapping, and the final content is rendered and streamed in real-time for VR interaction using Blender (OpenXR).

to generate both facial and body animations. TalkSHOW employs a VQ-VAE for body animation, while a standard encoder-decoder network predicts facial expressions. FaceFormer uses an autoregressive transformer (Vaswani, 2017) for facial expressions, and AMUSE employs a conditional latent diffusion model (Rombach et al., 2022). In summary, while EMAGE and TalkSHOW both use VQ-VAE, EMAGE leverages dual training paths (masked gesture recognition and audio-conditioned gesture generation with a switchable cross-attention layer) to effectively merge body hints and audio features and disentangle gesture decoding. In contrast, TalkSHOW trains face and body components separately, autoregressively predicting body and hand motion while incorporating facial expressions from the face decoder. Meanwhile, AMUSE is specially trained for emotional motion generation; since it focuses solely on emotional gesticulation without facial animation, we complement it with FaceFormer for full-body animation sequences.

For dialogue, we generate template responses to scenario-based questions. The text is then fed into a TTS model, which generates speech with appropriate intonation. These intonations drive the emotional arousal-related gestures produced by all models, ensuring alignment between speech and gestures. We use PlayHT TTS (PlayHT, 2025) to generate emotional speech given text inputs. For a given script, speech is generated with a storytelling narrative style for an adult male, featuring neutral tempo and loudness. Once the models have produced their outputs, GPU acceleration is used to render the meshes in Blender. We incorporate the body shape β parameter and import the *.npz* data into Blender through the SMPL-X addon (Pavlakos et al., 2019), which applies a sample albedo texture upon import, as shown in Figure 3-top.

3.2.1 Real human animation reconstruction

We also employ a video-based regression model to reconstruct animations from real actor gestures and expressions, allowing us

to compare the performance of synthetic animation against real human motion capture. The model processes a driving video of a real actor and outputs per-frame mesh objects. Specifically, we use PIXIE (Feng et al., 2021a) to estimate θ, ψ, and gender-specific shape β and α, while DECA (Feng et al., 2021b) extracts high-fidelity 3D facial displacements. For the reconstruction-based animation, we record an actor responding to scenario-based questions while another individual poses the questions. Video frames are extracted and processed by PIXIE and DECA to obtain geometry, α, and lighting information. The audio from the original video is used to synchronize lip movements with the spoken words. Detailed shaded textures, including 3D displacements, are applied by mapping UV textures onto the 3D body mesh on a per-frame basis. Each frame is then exported as a Wavefront OBJ file with shaded textures via PyTorch3D (Ravi et al., 2020). Finally, using Blender's Geometry Nodes editor, we generate instances of objects from a collection and place them on points derived from the mesh, animating the mesh sequences with the geometry node modifier, as shown in Figure 3-bottom. All animations share the same outdoor environment background. For inference, we use the default model hyperparameters provided by the original implementations of all methods: EMAGE, TalkSHOW, AMUSE, FaceFormer, PIXIE, and DECA. All input audio was sampled at 16 kHz. We used Blender 3.4 along with the built-in *VR Scene Inspection* add-on for VR streaming. The SMPL-X Blender add-on (v1.1) was used, along with the SMPL-X mesh, textures, and UV map (v1.1, NPZ+PKL format).

4 User study

4.1 Research questions

We address the following research questions for animations representing two emotional arousal categories:
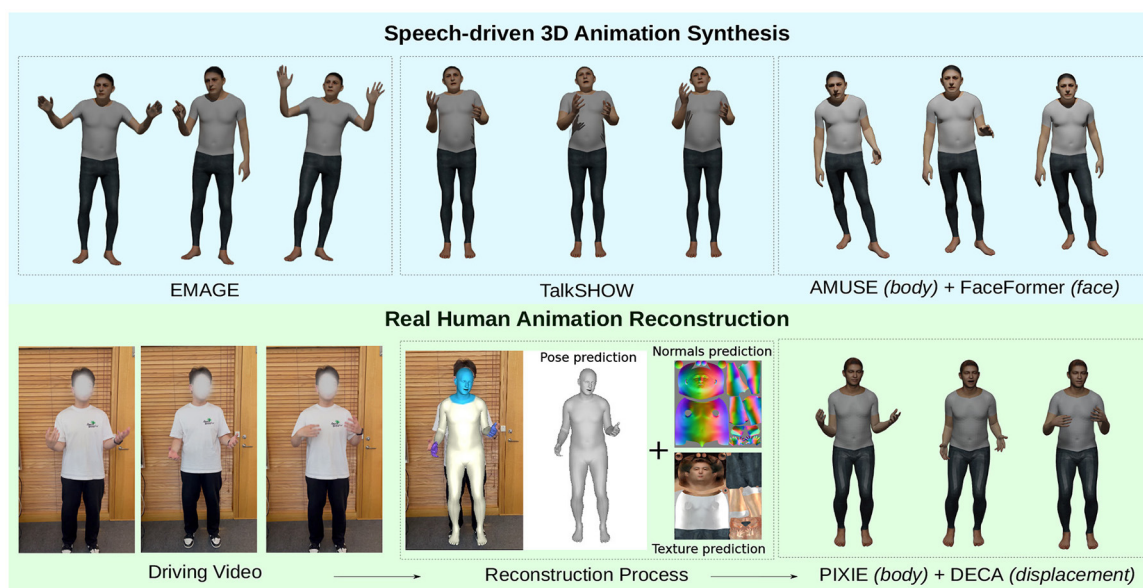
FIGURE 3
Qualitative evaluation. **Top**: Specific frames from the generated animation sequences using EMAGE (Liu et al., 2024), TalkSHOW (Yi et al., 2023), and a combination of AMUSE (body animation) (Chhatre et al., 2024) and FaceFormer (facial expressions) (Fan et al., 2021). **Bottom**: The workflow for generating reconstruction-based animations from real human facial expressions and body gestures using driving video input, which serves as our baseline. The reconstruction method PIXIE (Feng et al., 2021a) + DECA (Feng et al., 2021b) predicts pose parameters, normal maps, and textures, which are combined and rendered. Specific frames from the resulting video-based reconstruction animations are shown in the bottom right.

- **RQ1 (Perceived Animation Realism):** *"Which generative method demonstrates the highest perceived realism during a social interaction?"*
- **RQ2 (Perceived Animation Naturalness):** *"Which generative method demonstrates the highest naturalness in terms of facial expressions and bodily gestures?"*
- **RQ3 (Perceived Animation Enjoyment):** *"Do the methods influence the perceived level of enjoyment?"*
- **RQ4 (Perceived Interaction Quality):** *"Do the methods show differences in the quality of experienced interaction?"*
- **RQ5 (Perceived Animation Diversity):** *"Can participants perceive motion diversity between two virtual character animations of the same speech utterance with neutral emotion, presented side by side?"*
- **RQ6 (Perceived Animation Emotion):** *"Can participants correctly identify the arousal level in the generated animation that the model was given as input?"*

4.2 Participants

We recruited 48 participants (28 males, 20 females) aged 19-48 ($\mu = 26.71, SD = 5.30$) via internal channels at the local University. When asked about their recent experiences with virtual environments, 70.8% reported playing videogames in the past 12 months, and their previous enjoyment with VR experiences varied as follows: "below average" (6.25%), "average" (33.3%), "good" (37.5%), and "very good" (22.9%). All participants were recruited through an internal email system and received a gift card as

compensation. The study conformed to the Declaration of Helsinki and was approved by the local ethical committee.

4.3 Experiment conditions

We conducted a within-subject experiment with two independent variables: method (EMAGE, TalkSHOW, PIXIE+DECA, and AMUSE+FaceFormer) and scenario [Happy Emotion Animation (HEA), Neutral Emotion Animation (NEA), and Animation Diversity (DV)]. The HEA and NEA scenarios involve interactions with an agent displaying happiness and neutral animations, respectively. The DV scenario employs two different PyTorch noise seeds to generate distinct animations of two agents performing the same speech utterance with neutral emotion. In PyTorch, setting a fixed random seed helps control sources of randomness—allowing repeated executions on the same platform and device to produce identical outputs—, which lets us opt into deterministic implementations for certain operations. In the HEA and NEA scenarios, participants engage in one short conversation, whereas in the DV scenario they participate in two conversations. To systematically test method effects, we combined the four animation sources–three generative models and one based on a real human performance—with the three scenarios, yielding twelve experimental conditions. This design enables us to compare the effectiveness of the generative models in expressing emotionally expressive animation both between themselves and against the baseline (PIXIE+DECA) by measuring user perceptions during interaction with the virtual character. The ordering of conditions per participant was counterbalanced using a Latin Square design.

The scenario design follows principles from Fraser et al. (2022). Specific frames from all scenarios are shown in Figure 3.

4.3.1 Happy Emotion Animation (HEA)

Participants engage in a short conversation where the agent expresses happiness. The prompt is "*Past accomplishment*", and the agent responds with "*I engineered an AI-driven healthcare diagnostic tool, enhancing medical professionals' capabilities for rapid and accurate disease identification and treatment*", accompanied by consistent gestures and facial expressions generated by each method. This pre-generated response and motion are produced by the system described in Section 3.2 and configured to convey high arousal (happiness).

4.3.2 Neutral Emotion Animation (NEA)

This scenario mirrors the HEA condition, but conveying mid arousal instead (neutral emotion). The prompt is "*Way to relax*", and the response is "*I escape to a secluded garden, where the rustle of leaves and blooming flowers ease my mind*".

4.3.3 Animation Diversity (DV)

In this scenario, participants encounter two agents under the prompt "*Christmas plans*". Both agents respond with "*This Christmas, I'm eager to create handmade decorations and share the festive spirit with those around me*", each displaying motion-diverse body gestures and facial expressions generated from neutral emotion input, and presented side by side.

4.4 Survey

We used a 21-item questionnaire to gauge how each experimental condition influenced perception, social presence, and interaction quality. Three items collected demographics and prior VR exposure, while twelve items—split evenly between Happy and Neutral arousal blocks–assessed perceived realism [from the Networked Minds Social Presence Inventory (Biocca et al., 2003)], facial- and body-naturalness [adapted from Fraser et al. (2022)], interaction quality (Rogers et al., 2021), emotional arousal level (Biocca et al., 2003), and animation diversity (Conley et al., 2018; Cooperrider, 2020). Six additional post-study items, also adapted from Fraser et al. (2022), captured overall realism, interaction quality, face- and body-naturalness, diversity, and open-ended feedback. All conditions used five-point Likert items, except for perceived emotional arousal, which had three levels (high, medium, low), and the diversity item, which was a binary choice. Some prompts were slightly reworded to match the scope of our study. In Table 2 we provide a complete list of all questions, their primary sources, the subjective metrics they assess, and their intended applicability.

4.5 Apparatus

We used an HTC VIVE Pro 2 Head-Mounted Display (90 FPS, 120° FOV, 2448×2448 resolution per eye) with integrated

headphones. Two SteamVR 2.0 base stations tracked participants' positions. The virtual environment was created in Blender 3.4 with OpenXR-based SteamVR integration. The 30 FPS animation was played at 90 Hz in the VR headset using frame duplication, running on a desktop computer with an Intel i9-13900K CPU, 64GB RAM, and an NVIDIA RTX A6000 GPU. To ensure synchronized facial expressions and gestures despite method latency (section 6.4), speech and animations were pre-generated before the experiment and then streamed and rendered in real-time during user interaction.

4.6 Procedure

Participants received an introduction to the study and provided written consent. Once sat down and wearing the headset, they were greeted by a virtual character positioned $1.5\,m$ away, allowing them to position themselves comfortably for eye contact. We kept the interpersonal distance and outdoor scene constant, in an effort to eliminate confounding effects of proxemics and place illusion. They then removed the headset to complete a pre-experiment survey. Next, participants experienced the twelve conditions (four methods × three scenarios) in a counterbalanced order, one trial at a time. Before each trial, they were shown a paper with the conversation prompt and then wore the headset to interact with the virtual character. After each trial, they removed the headset to complete a condition-specific survey, before moving on to the next trial. Upon finishing all scenarios, participants completed a post-experiment survey.

4.7 Data analysis

Because the collected data did not satisfy the assumption of normality, we employed the aligned rank transform (ART), a non-parametric method suitable for factorial analyses (Wobbrock et al., 2011). Specifically, we used an ART ANOVA for all statistical tests and applied Bonferroni corrections for pairwise comparisons.

5 Results

5.1 Perceived animation realism

We found that the methods did not significantly influence realism: EMAGE ($Md = 2, IQR = 2$), TalkSHOW ($Md = 3, IQR = 2$), PIXIE+DECA ($Md = 3, IQR = 2$), and AMUSE+FaceFormer ($Md = 3, IQR = 2$). This result was confirmed by a non-significant main effect for methods [$F(3, 141) = 1.5, p = 0.2, \eta^2 = 0.03$]. However, we discovered that the happy emotion condition ($Md = 3, IQR = 2$) yielded higher realism ratings than the neutral emotion condition ($Md = 2.5, IQR = 2$), as supported by a statistically significant main effect for emotion [$F(1, 47) = 11.5, p < 0.001, \eta^2 = 0.2$]. Finally, no statistically significant interaction effect was observed for methods × emotion [$F(3, 141) = 1.6, p = 0.17, \eta^2 = 0.03$].

TABLE 2 Questions used in the perceptual study across VR conditions.

Cond.	Questions	Primary reference	Measures assessed	Intended applicability
Pre	Did you play any video game in the past 12 months?	Fraser et al., 2022	Exposure	Cross media
	How was your experience with virtual environments?	Fraser et al., 2022		
	Write positive and negative experiences for these interactions.	Fraser et al., 2022		
HEA	I felt like I was interacting with a real person.	Biocca et al., 2003	APR/ SP	XR
	Their facial expressions looked natural.	Fraser et al., 2022	ANF	Cross media
	Their body movements looked natural.	Fraser et al., 2022	ANB	Cross media
	Rate your enjoyment during the interaction.	Rogers et al., 2021	AE	Cross media
	Was the interaction warm and comfortable?	Rogers et al., 2021	IQ	Cross media
	What was your perceived emotion of the avatar's animation?	Biocca et al., 2003	EAR	Emotional interactions
NEA	I felt like I was interacting with a real person.	Biocca et al., 2003	APR/ SP	XR
	Their facial expressions looked natural.	Fraser et al., 2022	ANF	Cross media
	Their body movements looked natural.	Fraser et al., 2022	ANB	Cross media
	Rate your enjoyment during the interaction.	Rogers et al., 2021	AE	Cross media
	Was the interaction warm and comfortable?	Rogers et al., 2021	IQ	Cross media
	What was your perceived emotion of the avatar's animation?	Biocca et al., 2003	EAR	Emotional interactions
DV	Do you perceive diversity in the two animations presented?	Conley et al., 2018; Cooperrider, 2020	AD	Cross media
Post	Did you feel the strongest sense of closeness to your conversational partner?	Biocca et al., 2003	SP	XR
	Did the interaction with the avatar feel mostly like a real person?	Biocca et al., 2003	SP	XR
	The avatar's facial expressions looked the most natural/realistic.	Fraser et al., 2022	ANF	Cross media
	The avatar's body movements looked the most natural/realistic.	Fraser et al., 2022	ANB	Cross media
	Provide qualitative feedback about your overall VR interaction experience.	Fraser et al., 2022	Feedback	Cross media

The table lists each condition (Cond.), the corresponding survey questions, the primary source of the item, and the construct it measures. Abbreviations for the measures: APR, Animation-Perceived Realism; ANF, Animation-Naturalness of Facial Expressions; ANB, Animation-Naturalness of Body Gestures; AE, Animation Enjoyment; IQ, Interaction Quality; EAR, Animation Emotional Arousal Recognition; AD, Perceived Animation Diversity; SP, Social Presence; OF, Open-ended Feedback; XR, Extended Reality (VR/AR/MR). Cross media represents any media platform—video, XR, or other formats.

5.2 Perceived animation naturalness of facial expressions

PIXIE+DECA ($Md = 3, IQR = 2$) resulted in higher ratings for the naturalness of facial expressions compared to EMAGE ($Md = 2, IQR = 1$), TalkSHOW ($Md = 3, IQR = 2$), and FaceFormer ($Md = 2, IQR = 1$). This was confirmed by a statistically significant main effect for methods [$F(3, 141) = 3.3, p = 0.02, \eta^2 = 0.07$]. Pairwise comparisons showed significant differences between EMAGE and PIXIE+DECA ($p = 0.01$), but not among the other pairs ($p > 0.05$). No significant differences were found between the happy emotion ($Md = 3, IQR = 1$) and neutral emotion ($Md = 2, IQR = 1$) conditions [$F(1, 47) = 1.49, p = 0.22, \eta^2 = 0.03$]. However, a statistically significant interaction effect for methods × emotion was observed [$F(3, 141) = 4.1, p = 0.007, \eta^2 = 0.08$]. Pairwise comparisons revealed significant differences between the neutral emotion condition in EMAGE and PIXIE+DECA ($p = 0.01$), and between TalkSHOW happy emotion and EMAGE neutral

emotion ($p = 0.0238$); the remaining comparisons were not significant ($p > 0.05$).

5.3 Perceived animation naturalness of body gestures

We found that methods did not significantly affect the naturalness of bodily movements: EMAGE ($Md = 3, IQR = 2$), TalkSHOW ($Md = 3, IQR = 2$), PIXIE ($Md = 3, IQR = 2$), and AMUSE ($Md = 3, IQR = 2$). This was confirmed by a non-significant main effect for methods [$F(3, 141) = 1.3, p = 0.26, \eta^2 = 0.03$]. However, the happy emotion condition ($Md = 3, IQR = 2$) resulted in higher naturalness ratings than the neutral emotion condition ($Md = 3, IQR = 2$), a difference supported by a statistically significant main effect for emotion [$F(1, 47) = 6.4, p = 0.01, \eta^2 = 0.12$]. No significant interaction effect was observed for methods × emotion [$F(1, 141) = 1.57, p = 0.19, \eta^2 = 0.03$].
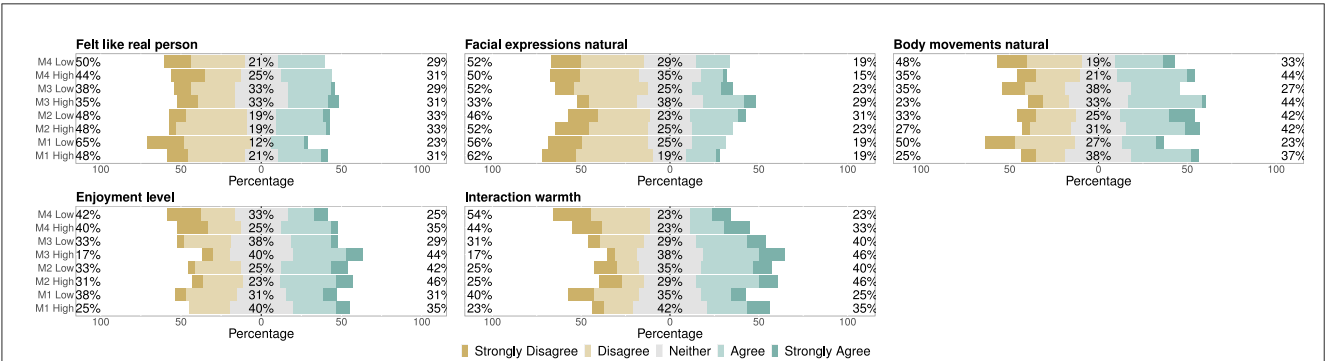
FIGURE 4
Summary of likert scale results. Summary of Likert scale ratings for Animation Realism (avatar felt like a real person), Animation Naturalness (facial expressions; body movements), Animation Enjoyment, and Interaction Quality (interaction warmth). For brevity, we denote EMAGE, TalkSHOW, PIXIE+DECA, and AMUSE+FaceFormer as M1, M2, M3, and M4, respectively, and use "High" and "Low" to represent happy and neutral emotions.

5.4 Perceived animation enjoyment

We found that the methods did not significantly influence enjoyment levels: EMAGE ($Md = 3, IQR = 2$), TalkSHOW ($Md = 3, IQR = 2$), PIXIE+DECA ($Md = 3, IQR = 1.25$), and AMUSE+FaceFormer ($Md = 3, IQR = 2$). Similarly, there was no significant difference between the happy emotion ($Md = 3, IQR = 2$) and neutral emotion ($Md = 3, IQR = 2$) conditions. These findings were supported by non-significant main effects for methods [$F(3, 141) = 2.4, p = 0.06, \eta^2 = 0.05$] and emotion [$F(1, 47) = 2.6, p = 0.11, \eta^2 = 0.05$]. Additionally, no statistically significant interaction effect was found for methods × emotion [$F(3, 141) = 1.05, p = 0.36, \eta^2 = 0.022$].

5.5 Perceived interaction quality

TalkSHOW ($Md = 3, IQR = 1.25$) resulted in higher ratings for interaction quality compared to EMAGE ($Md = 2, IQR = 1$), PIXIE+DECA ($Md = 3, IQR = 1$), and AMUSE+FaceFormer ($Md = 3, IQR = 2$). This difference was supported by a statistically significant main effect for methods [$F(3, 141) = 4.2, p < 0.01, \eta^2 = 0.08$]. Pairwise comparisons indicated significant differences between TalkSHOW and AMUSE+FaceFormer ($p = 0.027$), while the other comparisons were not significant ($p > 0.05$). No significant differences were observed between the happy ($Md = 3, IQR = 2$) and neutral ($Md = 3, IQR = 2$) emotion conditions [$F(1, 47) = 4, p = 0.051, \eta^2 = 0.07$]. Furthermore, no significant interaction effect was found for methods × emotion [$F(3, 141) = 1.57, p = 0.2, \eta^2 = 0.03$]. A summary of the Likert scale results for realism, facial expressions, bodily movements, enjoyment, and interaction quality is shown in Figure 4.

5.6 Animation emotional arousal recognition

As the last question in the six-item survey, participants rated the animations arousal for both HEA and NEA conditions. After being

TABLE 3 Arousal recognition rates by method and sequence.

Method	Seq.	High	Mid	Low
EMAGE	H	55.5	35.2	9.3
	N	10.8	72.2	17
TalkSHOW	H	56	34.8	9.2
	N	5.8	78.4	15.8
PIXIE+DECA	H	61.5	30.4	8.1
	N	2.2	89.58	8.2
AMUSE+FaceFormer	H	70.83	23.1	6.1
	N	6.8	74.4	18.8

For each animation method and sequence (H = high-arousal for happiness, N = mid-arousal for neutral), participants' percent correct identification in the three response categories.

told to judge the perceived emotional arousal of each clip, they chose one of three options: high, medium, or low arousal. Overall, participants correctly identified high-arousal clips 60.94% of the time and mid-arousal clips 78.65% of the time. By method, EMAGE had a recognition percentage of 55.5% on high and 72.2% on mid, TalkSHOW 56.0% and 78.4%, PIXIE + DECA 61.5% and 89.58%, and AMUSE + FaceFormer 70.83% and 74.4%, respectively. Thus, AMUSE + FaceFormer led in high-arousal recognition, while PIXIE + DECA excelled at mid-arousal detection. The detailed confusion matrix for two stimulus levels (high and mid arousal) across three response options (high, mid, low) is shown in Table 3. Correct identifications are highlighted in blue, while any confusions in which a high- or mid-arousal stimulus was classified as low arousal are shaded in violet.

To further analyze arousal recognition, we used a deep learning-based motion extractor (Petrovich et al., 2021; Chhatre et al., 2024) trained on motion capture data to predict one of eight emotion classes (an extended Ekman-style eight-class taxonomy: neutral, happy, angry, sad, contempt, surprise, fear, disgust). We present the predicted emotion recognition probabilities in Table 4, where the best-performing methods' happy and neutral sequences are highlighted in blue and second best in yellow, while the emotions with which the method is confused are highlighted in violet.

TABLE 4 Emotion recognition accuracy for happy and neutral animations.

Method	Seq.	N	H	A	S	C	Su	F	D
EMAGE	H	8.5	19.5	9.2	30.2	7.8	8.1	8.0	8.7
	N	20.0	8.2	8.5	31.0	9.0	8.0	8.0	7.3
TalkSHOW	H	17.5	28.0	10.8	8.3	9.6	8.2	10.5	7.1
	N	30.0	8.5	8.2	8.8	9.1	9.0	18.0	8.4
PIXIE+DECA	H	6.8	44.0	18.7	6.4	6.6	6.5	5.2	5.8
	N	44.2	6.9	5.4	8.0	5.3	6.5	16.8	6.9
AMUSE+FaceFormer	H	3.2	56.0	3.6	3.5	3.7	18.9	5.5	5.6
	N	54.3	5.0	3.5	19.0	4.8	2.4	5.7	5.3

Each row shows the softmax probability distribution (%) over the eight Ekman emotions for a given method and sequence. N, H, A, S, C, Su, F, and D represent neutral, happiness, anger, sadness, contempt, surprise, fear, and disgust, respectively. See Section 6.2 for discussion.

5.7 Animation diversity

Participants were asked to judge whether two side-by-side virtual-character animations–generated from distinct initial conditions as described in Section 4.3–appeared diverse. Because this diversity item was a binary choice, we did not subject it to statistical analysis. AMUSE+FaceFormer was rated most effective, with 95.8% of participants perceiving diversity. In contrast, EMAGE received the lowest ratings, with 70.8% reporting perceived diversity and 18.8% indicating no diversity. Both TalkSHOW and PIXIE+DECA received 79.2% of participants reporting perceived diversity. To complement our statistical analysis, we computed the Euclidean distance (2-norm) between joints on the SMPL-X axis angles, yielding diversity scores of 2.5336 for EMAGE, 2.0777 for TalkSHOW, and 2.9360 for AMUSE+FaceFormer; PIXIE+DECA shows no diversity due to its deterministic reconstruction approach.

5.8 Post-experiment feedback: perceived closeness and realism

Participants evaluated their experiences using 5-point Likert scale responses regarding closeness, perceived realism, and the naturalness of facial expressions and bodily movements. The post-study items were collected only once per participant–after all methods had been experienced–we treat these four measures as overall user impressions rather than method-specific comparisons. Accordingly, we report only descriptive statistics and do not perform factorial tests. Post-study ratings yielded a median sense of closeness ($Md = 2, IQR = 1$), agent realism ($Md = 3, IQR = 2$), facial-expression naturalness ($Md = 3, IQR = 1$), and body-gesture naturalness ($Md = 3, IQR = 2$), indicating an overall mildly positive perception of the virtual characters social presence and animation quality. EMAGE and TalkSHOW received the lowest ratings, with 28 and 30 participants, respectively, rating closeness as "A little". PIXIE+DECA performed best, with 24 participants reporting "Quite a bit" of closeness and 27 finding the agent realistic. PIXIE+DECA also scored highest for natural facial expressions, with 23 participants rating them as "Quite a bit". AMUSE+FaceFormer received more balanced feedback, with 22 participants finding the agent realistic and 23 rating the bodily movements as natural.

6 Discussion

6.1 Scenario design

Our scenarios focus on everyday conversations, each associated with an internal emotional state. For example, a relaxation topic corresponds to a neutral emotional stance, while a past achievements topic evokes a happy emotional state. This setup explores how varying emotional cues affect behavior and perception. In a "passion" scenario, audio and gestures convey energetic or happy expressions, whereas in a "relaxation" scenario, they are more subdued. We then evaluate the extent to which the methods can generate distinguishable emotional levels. Regarding animation diversity, we measure how much gesture variation is acceptable before the virtual characters identity appears inconsistent for the same speech utterance, as if a participant was interacting with a different entity within the same scenario.

6.2 Emotional 3D animation

Our findings show that the emotion category significantly affects animation realism. Happy emotion animations with energetic gestures were perceived as more realistic than neutral emotion animations, indicating that high-arousal, happy expressions have a stronger social presence during an interaction across all methods (RQ1). In terms of facial expression naturalness, PIXIE+DECA outperformed the other methods–especially in neutral emotion scenarios—demonstrating a superior ability to capture subtle facial cues. Additionally, emotional interaction revealed that PIXIE+DECA consistently performed better (particularly compared to EMAGE in neutral conditions), primarily due to DECA's robust capture of 3D facial displacements, which enhances the base expressions predicted by PIXIE; no speech-driven face animation method is comparable to the reconstruction-based real human facial expressions compatible with SMPL-X meshes (RQ2-face). For body movement naturalness, emotion again played a key role: happy emotion movements were rated more natural (RQ2-body), mirroring the results for animation realism (RQ1). While enjoyment levels were similar across all methods (RQ3), TalkSHOW outperformed the others in interaction quality–especially when compared to EMAGE and

AMUSE+FaceFormer—suggesting that TalkSHOWs output may support a stronger interactive connection with users (RQ4).

Survey data shows that 60.94% of participants correctly identified the happy emotion condition, while 78.65% correctly recognized the neutral emotion condition, indicating that mid-arousal gestures were easier to identify (RQ6). PIXIE+DECA achieved the highest accuracy (89.58%) for neutral emotion, whereas AMUSE+FaceFormer performed best for happy emotion (70.83%), demonstrating that AMUSE+FaceFormer animations are easier to recognize for happy emotion compared to other methods. EMAGE exhibited balanced accuracy for both conditions, while TalkSHOW and PIXIE+DECA showed a trend toward more accurate mid-arousal identification.

The effectiveness of each model in generating distinguishable emotional levels depends on its architecture and processing approach (see Table 1). All methods are generative and probabilistic, but differ in their preprocessing approaches; EMAGE and AMUSE include unique processing steps, whereas TalkSHOW uses standard inputs without specialized preprocessing.

6.2.1 EMAGE and AMUSE

Extract disentangled latent representations for speech content, emotion, style, and rhythm. These robust representations allow for better alignment with arousal cues rather than merely producing varied animations.

6.2.2 PIXIE

Operates purely on video input, reconstructing realistic animation directly from a human actors performance. Although this can yield high-quality results, it relies on the actors expressiveness and does not create new gestures.

Using this statistical deep learning emotion recognition metric for motion sequences, we observe that AMUSE+FaceFormer and PIXIE+DECA demonstrate the highest emotion recognition scores, with AMUSE+FaceFormer achieving 56% accuracy for happy emotion and 54.3% for neutral emotion. Specifically, AMUSE+FaceFormer predictions confused Happy with Surprise and Neutral with Sad; PIXIE+DECA confused Happy with Angry and Neutral with Fear. In contrast, TalkSHOW and EMAGE demonstrate lower emotion recognition accuracy, with TalkSHOW confusing Happy with Neutral and Neutral with Fear, and EMAGE showing high confusion with Sad in both sequences. These quantitative findings align with our user study data, indicating that PIXIE+DECA excels at capturing high-quality animations, although depending on the actors performance, whereas audio-based methods can independently generate synthetic animations with disentangled emotion and content, producing more clearly distinguishable gesture arousal–with the speech-driven method AMUSE+FaceFormer showing the highest accuracy.

6.3 Animation diversity

The perceived animation diversity varied significantly across models, with AMUSE+FaceFormer standing out—95.8% of participants noticed diverse animations. In contrast, EMAGE scored lowest at 70.8%, while 79.2% of participants observed diversity with the other models. Animation diversity is essential for crowd animations and extended interactions, where varied gestures, movements, and contexts create engaging, lifelike experiences; it also applies to both speech-driven and idle animations, which are key to maintaining natural behavior in virtual characters. Additionally, quantitative measurements of animation diversity—computed as the Euclidean (2-norm) distance between SMPL-X joint axis angles—show that AMUSE+FaceFormer has the highest 2-norm, followed by EMAGE and TalkSHOW. These findings reinforce our statistical analysis, confirming that greater animation diversity enhances perceived interaction quality and supports RQ5.

6.4 Inference times

The inference times for producing 10-second animation sequences were as follows: EMAGE required 0.827s; TalkSHOW, 20.29s; PIXIE+DECA, 412.63s; and AMUSE+FaceFormer totaled 8.561s (2.557s for body animation, plus 5.337s for face animation). EMAGE is the fastest, making it particularly efficient for real-time or near real-time applications. AMUSE+FaceFormer strikes a balance between speed and complexity, being faster than TalkSHOW but slightly slower than EMAGE, while PIXIE+DECA is by far the slowest due to the complexity of video-based animation reconstruction.

6.5 Design recommendations

In our evaluation, we compared state-of-the-art speech-driven 3D emotional animation generation methods to examine their strengths and weaknesses, as well as how they shape user perception in VR. By comparing these generative approaches with the reconstruction of a real actor, we also investigated how closely current methods can replicate real human body and facial expressions. We note that marker-based motion capture yields higher-quality real-actor motion than our reconstructed animations. Based on our user study results, we note the following design recommendations.

6.5.1 Emotional modeling

While speech-to-animation methods often focus on lip-sync and body gestures, explicit emotion modeling is frequently overlooked. As shown in Table 4, all animation methods (EMAGE, TalkSHOW, AMUSE, FaceFormer) have considerable scope for improvement in emotion recognition (RQ6). Although AMUSE, which is trained on explicit audio emotion and person identity modeling for gesture generation, shows the best relative accuracy (56.0% for happiness and 54.3% for neutral), there remains a gap. Emotions are often confused with other emotional animation sequences across models. AMUSE achieves emotion modeling by disentangling driving speech into content, emotion, and style; however, its approach is limited to a single categorical emotion per sequence. Exploring multiple concurrent emotions in an animation

sequence is a promising direction for future research. Finally, scenario context may affect emotion perception: participants inferred emotions not only from gestures but also from the spoken content or from how convincingly the actors performance was rendered (for the video reconstruction method).

6.5.2 Animation generation for emotional states with lower arousal

In terms of animation realism (RQ1) and naturalness (RQ2 for both body and face), we observed that animations representing high-arousal, happy emotions consistently received higher ratings than those for neutral, low-arousal states. The generative models generally perform better when generating pronounced expressions compared to subtle, idle movements. This is likely because the motion datasets used for training typically include expressive sequences rather than calm, idle ones. Incorporating mocap datasets focused on calm or idle motions, such as breathing-based movement, could help models generalize to less exaggerated animations.

6.5.3 Joint modeling of facial expressions and body gestures

Among the evaluated methods, EMAGE jointly trains face and body, whereas TalkSHOW trains them separately within the same framework, and AMUSE does not address facial expressions. Even with joint training in EMAGE, no method achieved high ratings for facial expression naturalness (RQ2-face). This highlights the challenge of simultaneously learning both expression parameters and body gestures, largely due to the differences in data representations (face data uses 100 SMPL-X expression parameters, while body data is based on joint rotations in the world coordinate system). More robust data preprocessing and unified parameterization are needed to effectively train a single model for both full-body and facial animations.

6.5.4 Dyadic interaction feedback

All generative methods exhibited similarly low performance in animation enjoyment (RQ3) and interaction quality (RQ4), although TalkSHOW showed relatively higher interaction quality. This suggests that current evaluations, which often rely solely on statistical metrics, do not fully capture the user-centric experience in immersive interaction settings. Incorporating user-centered evaluation into the feedback mechanism is crucial, as it ensures that the generated animations effectively convey the intended emotion and enhance user enjoyment and interaction.

7 Limitations and future work

Our evaluation represents a useful first step but has several limitations that suggest valid directions for future research–namely, addressing latency issues, exploring the full spectrum of emotion categories, varying VR hardware setups, incorporating additional behavioral measures, and benchmarking against video-based reconstruction–all of which are detailed below.

7.1 Latency and turn-taking

Our evaluation employs a modular approach using several large deep generative models. For instance, AMUSE—a latent diffusion model with 440 million parameters—generates temporal SMPL-X motion parameters but suffers from slow inference due to extensive denoising steps and high GPU memory requirements (e.g., RTX A6000 with 48 GB). Similarly, the face generation models in TalkSHOW and FaceFormer, which are based on Transformer autoregressive architectures, experience longer inference times due to their sequential design. As these latencies limit real-time applications, we pre-generate speech and animations and stream them in real time for single-turn conversations, as noted in Section 4.5. Supporting multi-turn conversations, however, would require a fully real-time setup with no pre-generation, where the speech responses and corresponding animations must be generated and streamed simultaneously in VR. As noted in Section 6.4, EMAGE is currently the most suitable method for real-time interaction, achieving the lowest latency at 0.827s. To account for this, we introduce a fixed 5-second idle movement period between turns, where the agent adopts a neutral, forward-facing stance. In a fully real-time multi-turn conversation, idle motion and wait times would need to dynamically adapt to the length of the participant's response. To the best of our knowledge, no currently available system can generate idle body motion and trigger speech and animation generation in real time upon detecting the end of a users reply. Achieving this would require integrating speech-to-speech models with real-time animation generation, followed by synchronized playback of speech and gestures–an avenue we identify as future work. Such a system would also require careful consideration of hardware, as computing demands are expected to remain high.

7.2 Emotion categories

Emotions can be described as discrete classes or as points in a continuous affective space (see Section 2.1.1). To keep the study tractable, we restricted our experiment to two conditions–Happy (high arousal) and Neutral (mid arousal)–which provided clear initial insights while avoiding an exponential growth in the number of possible condition combinations. Extending the protocol to cover additional emotions across the full arousal-valence spectrum of the circumplex model is an important goal for future work.

7.3 VR apparatus

VR streaming and rendering are compute-intensive and heavily hardware-dependent. As described in Section 4.5, we mitigated these challenges by splitting the interaction into two phases: real-time streaming for interaction and precomputed full-body animation rendering, with other components generated in advance. Future improvements in VR hardware for rendering and streaming will further alleviate these limitations.

7.4 Video-based reconstruction

Although our video-based reconstruction method shows promising per-frame quality, its temporal coherence is lower. Frame-by-frame pose estimation, when played back at 30 FPS, leads to jittery animations (see Supplementary Video). Despite our initial expectation that video-based reconstruction of the real animation would yield the best performance across enjoyment (RQ3) and interaction quality (RQ4), our user study revealed that reconstruction methods did not excel in these areas, even though facial animation (RQ2-face) was enhanced by DECA-based face displacements. Future studies should explore reconstruction methods that improve temporal coherence and pose estimation for smoother animations at the desired frame rate.

7.5 Additional behavioral measures

Research on human-like behavior in virtual agents remains in its early stages. Our study is a useful first step, but a richer evaluation is needed—particularly on metrics such as eye-gaze patterns and task-completion time. Concepts central to believability and presence—including co-presence, plausibility, place illusion, the uncanny valley for interactive agents, and both subjective and inter-subjective symmetry—also require analysis. In addition, more complex social dynamics (e.g., group interaction and contact behavior such as self-contact, interpersonal contact, and ground contact) should be examined. Progress will depend on developing stronger generative models and testing them in more sophisticated realistic environments.

8 Conclusion

We present an evaluation of generative models for emotional 3D animation within an immersive VR environment, focusing on user-centric metrics–emotional arousal realism, naturalness, enjoyment, diversity, and interaction quality–in a real-time human-virtual character interaction scenario through a user study ($N = 48$). In this study, we systematically examined perceived emotional quality across three state-of-the-art speech-driven 3D animation methods and compared them to a real human reconstruction-based animation under two emotional conditions: happiness (high arousal) and neutral (mid arousal). Participants recognized emotions more accurately for generative methods that explicitly modeled animation emotions. User study data showed that generative models performed well for high-arousal emotion but struggled with subtle arousal emotion. Although reconstruction-based animations received higher ratings for facial expression quality, all generative methods exhibited lower ratings for animation enjoyment and interaction quality, highlighting the importance of incorporating user-centric evaluations into generative animation model development. All methods demonstrated acceptable animation diversity; however, differing inference times among generative methods, along with VR rendering latency, posed limitations. Lastly, while the

video-based reconstruction method (compatible with SMPL-X meshes) produced high-quality frame-level animations from driving videos, it lacked temporal coherence, leading to suboptimal performance in user ratings of animation enjoyment and interaction quality. Overall, these findings highlight the importance of integrating user-centric evaluations into the development of generative models to produce virtual animated agents that outperform rule-based and teleoperated techniques. Hence, we believe that evaluating models solely on technical metrics during development is insufficient to ensure that the animations convey the perceptual details we want end users to experience in conversational scenarios.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Robin Roy, KTH Public Information Request Coordinator. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

KC: Software, Data curation, Writing – original draft, Methodology, Formal analysis, Visualization, Resources, Investigation, Supervision, Conceptualization, Project administration, Validation, Writing – review & editing. RG: Validation, Writing – original draft, Writing – review & editing, Supervision, Investigation, Conceptualization. AM: Visualization, Validation, Investigation, Conceptualization, Writing – review & editing, Formal analysis, Writing – original draft. CP: Supervision, Writing – original draft, Conceptualization, Funding acquisition, Writing – review & editing.

Tairan Yin for insightful discussions, proofreading, and valuable feedback.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2025.1598099/full#supplementary-material

References

Meet Your Soul Machines AI Assistants. Available online at: https://www.soulmachines.com/soul-machines-studio (Accessed July 19, 2024).

Alexanderson, S., Nagy, R., Beskow, J., and Henter, G. E. (2023). Listen, denoise, action! Audio-driven motion synthesis with diffusion models. *ACM Trans. Graph.* 42, 1–20. doi: 10.1145/3592458

Andric, M., Solodkin, A., Buccino, G., Goldin-Meadow, S., Rizzolatti, G., and Small, S. L. (2013). Brain function overlaps when people observe emblems, speech, and grasping. *Neuropsychologia* 51, 1619–1629. doi: 10.1016/j.neuropsychologia.2013.03.022

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, eds. H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (New York, NY: Curran Associates, Inc.), 12449–12460.

Barreda-Ángeles, M., and Hartmann, T. (2021). Psychological benefits of using social virtual reality platforms during the covid-19 pandemic: the role of social and spatial presence. *Comput. Human Behav.* 127:107047. doi: 10.1016/j.chb.2021.107047

Bernardis, P., and Gentilucci, M. (2006). Speech and gesture share the same communication system. *Neuropsychologia* 44, 178–190. doi: 10.1016/j.neuropsychologia.2005.05.007

Biocca, F., Harms, C., and Burgoon, J. K. (2003). Toward a more robust theory and measure of social presence: review and suggested criteria. *Presence: Teleoperators and Virtual Environments* 12, 456–480. doi: 10.1162/105474603322761270

Bolkart, T. (2013). *Tf flame: Tensorflow Framework for the Flame 3d Head Model.* Available online at: https://github.com/TimoBolkart/TF_FLAME (Accessed July 21, 2025).

Bradley, M. M., Greenwald, M. K., Petry, M. C., and Lang, P. J. (1992). Remembering pictures: pleasure and arousal in memory. *J. Exp. Psychol. Learn. Memory Cognit.* 18, 379–90.

Brandstätter, K., and Steed, A. (2023). "Dialogues for one: Single-user content creation using immersive record and replay," in *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology* (New York, NY: Association for Computing Machinery).

Casanova, E., Weber, J., Shulby, C., Júnior, A. C., Gölge, E., and Ponti, M. A. (2021). YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. *arXiv* [preprint] arXiv:2112.02418. doi: 10.48550/arXiv.2112.02418

Cassell, J. (2000). Embodied conversational interface agents. *Commun. ACM* 43, 70–78. doi: 10.1145/332051.332075

Cassell, J., Vilhjálmsson, H. H., and Bickmore, T. W. (2001). "Beat: the behavior expression animation toolkit," in *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY: Association for Computing Machinery).

Chatziagapi, A., Morency, L.-P., Gong, H., Zollhöfer, M., Samaras, D., and Richard, A. (2025). Av-flow: Transforming text to audio-visual human-like interactions. *arXiv* [preprint] arXiv:2502.13133. doi: 10.48550/arXiv.2502.13133

Chhatre, K., Daněček, R., Athanasiou, N., Becherini, G., Peters, C., Black, M. J., et al. (2024). "AMUSE: Emotional speech-driven 3D body animation via disentangled latent diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA: IEEE Institute of Electrical and Electronics Engineers), 1942–1953. doi: 10.1109/CVPR52733.2024.00190

Chhatre, K., Guarese, R., Matviienko, A., and Peters, C. (2025). "Evaluating speech and video models for face-body congruence," in *Companion Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games* (New York, NY, USA: Association for Computing Machinery).

Conley, M. I., Dellarco, D. V., Rubien-Thomas, E., Cohen, A. O., Cervera, A., Tottenham, N., et al. (2018). The racially diverse affective expression (radiate) face stimulus set. *Psychiatry Res.* 270, 1059–1067. doi: 10.1016/j.psychres.2018.04.066

Convai (2025). *Embodied ai Characters for Virtual Worlds*. Available online at: https://convai.com/ (Accessed July 19, 2024).

Cooperrider, K. (2020). Universals and diversity in gesture: research past, present, and future. *Gesture* 18:5432c. doi: 10.31234/osf.io/5432c

Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., and Black, M. J. (2019). "Capture, learning, and synthesis of 3d speaking styles," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019* (Long Beach, CA: Computer Vision Foundation / IEEE), 10101–10111.

Daněček, R., Chhatre, K., Tripathi, S., Wen, Y., Black, M., and Bolkart, T. (2023a). *Emotional Speech-Driven Animation With Content-Emotion Disentanglement*. New York City: ACM.

Daněček, R., Chhatre, K., Tripathi, S., Wen, Y., Black, M. J., and Bolkart, T. (2023b). *Emotional Speech-Driven Animation With Content-Emotion Disentanglement* (New York, NY: Association for Computing Machinery).

De Stefani, E., and De Marco, D. (2019). Language, gesture, and emotional communication: an embodied view of social interaction. *Front. Psychol.* 10:2063. doi: 10.3389/fpsyg.2019.02063

Deichler, A., Beskow, J., and Werner, A. W. (2024). "Gesture evaluation in virtual reality," in *GENEA: Generation and Evaluation of Non-verbal behavior for Embodied Agents Workshop 2024* (New York, NY: Association for Computing Machinery).

Della Greca, A., Ilaria, A., Tucci, C., Frugieri, N., and Tortora, G. (2024). "A user study on the relationship between empathy and facial-based emotion simulation in virtual reality," in *Proceedings of the 2024 International Conference on Advanced Visual Interfaces* (New York, NY, USA: Association for Computing Machinery).

Di Natale, A. F., Simonetti, M. E., La Rocca, S., and Bricolo, E. (2023). Uncanny valley effect: a qualitative synthesis of empirical research to assess the suitability of using virtual faces in psychological research. *Comp. Human Behav. Reports* 10:100288. doi: 10.1016/j.chbr.2023.100288

Dosovitskiy, A. (2020). An image is worth 16x16 words: transformers for image recognition at scale. *arXiv* preprint arXiv:2010.11929. doi: 10.48550/arXiv.2010.11929

Ekman, P. (1993). Facial expression and emotion. *Am. Psychol.* 48, 384–392. doi: 10.1037//0003-066X.48.4.384

Fan, Y., Lin, Z., Saito, J., Wang, W., and Komura, T. (2021). Faceformer: Speech-driven 3d facial animation with transformers. *arXiv* [preprint] arXiv:2112.05329. doi: 10.1109/CVPR52688.2022.01821

Fan, Y., Lin, Z., Saito, J., Wang, W., and Komura, T. (2022). "Faceformer: Speech-driven 3d facial animation with transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022* (New Orleans, LA: IEEE), 18749–18758.

Feng, Y., Choutas, V., Bolkart, T., Tzionas, D., and Black, M. J. (2021a). "Collaborative regression of expressive bodies using moderation," in *International*

Conference on 3D Vision (3DV) (London: IEEE Institute of Electrical and Electronics Engineers).

Feng, Y., Feng, H., Black, M. J., and Bolkart, T. (2021b). Learning an animatable detailed 3D face model from in-the-wild images. *ACM Trans. Graph.* 40, 1–88. doi: 10.1145/3450626.3459936

Fraser, A., Branson, I., Hollett, R., Speelman, C., and Rogers, S. (2022). Expressiveness of real-time motion captured avatars influences perceived animation realism and perceived quality of social interaction in virtual reality. *Front. Virtual Reality* 3:981400. doi: 10.3389/frvir.2022.981400

Gentilucci, M., Bernardis, P., Crisi, G., and Volta, R. D. (2006). Repetitive transcranial magnetic stimulation of broca's area affects verbal responses to gesture observation. *J. Cogn. Neurosci.* 18, 1059–1074. doi: 10.1162/jocn.2006.18.7.1059

Ghorbani, S., Ferstl, Y., Holden, D., Troje, N. F., and Carbonneau, M.-A. (2023). Zeroeggs: Zero-shot example-based gesture generation from speech. *Comp. Graphics Forum* 42, 206–216. doi: 10.1111/cgf.14734

Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., and Malik, J. (2019). "Learning individual styles of conversational gesture," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Long Beach: IEEE), 3497–3506.

Guadagno, R. E., Blascovich, J., Bailenson, J. N., and McCall, C. A. (2007). Virtual humans and persuasion: The effects of agency and behavioral realism. *Media Psychol.* 10, 1–22. Available online at: https://www.tandfonline.com/doi/full/10.1080/15213260701300865

Gunter, T. C., and Bach, P. (2004). Communicating hands: Erps elicited by meaningful symbolic hand postures. *Neurosci. Lett.* 372, 52–56. doi: 10.1016/j.neulet.2004.09.011

Habibie, I., Elgharib, M. A., Sarkar, K., Abdullah, A., Nyatsanga, S. L., Neff, M., et al. (2022). "A motion matching-based framework for controllable gesture synthesis from speech," in *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)* (New York, NY: Association for Computing Machinery).

Hancock, J. T., Landrigan, C., and Silver, C. (2007). "Expressing emotion in text-based communication," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '07* (New York, NY: Association for Computing Machinery), 929–932.

He, Y., Steines, M., Sommer, J., Gebhardt, H., Nagels, A., Sammer, G., et al. (2018). Spatial-temporal dynamics of gesture-speech integration: a simultaneous eeg-fmri study. *Brain Struct. Funct.* 223, 3073–3089. doi: 10.1007/s00429-018-1674-5

Hedra (2025). *Video for Everyone.* Available online at: https://www.hedra.com/ (Accessed July 19, 2024).

Heesen, R., Szenteczki, M. A., Kim, Y., Kret, M. E., Atkinson, A. P., Upton, Z., et al. (2024). Impact of social context on human facial and gestural emotion expressions. *iScience* 27:110663. doi: 10.1016/j.isci.2024.110663

Herrera, F., Oh, S. Y., and Bailenson, J. N. (2018). Effect of behavioral realism on social interactions inside collaborative virtual environments. *Presence: Virtual Augment. Reality* 27, 163–182. doi: 10.1162/pres_a_00324

Huang, X., and Lajoie, S. P. (2023). Social emotional interaction in collaborative learning: why it matters and how can we measure it? *Social Sci Humanit Open.* 7:100447. doi: 10.1016/j.ssaho.2023.100447

Huesser, C., Schubiger, S., and Çöltekin, A. (2021). "Gesture interaction in virtual reality: A low-cost machine learning system and a qualitative assessment of effectiveness of selected gestures vs. gaze and controller interaction," in *Human-Computer Interaction INTERACT 2021: 18th IFIP TC 13 International Conference* (Berlin: Springer-Verlag), 151–160.

Inworld (2025). *The Leading AI Engine for Games.* Available online at: https://inworld.ai/ (Accessed July 19, 2024).

Karras, T., Aila, T., Laine, S., Herva, A., and Lehtinen, J. (2017). Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graph. (TOG)* 36, 1–12. doi: 10.1145/3072959.3073658

Kim, J., Kong, J., and Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *arXiv.* doi: 10.48550/arXiv.2106.06103

Kokkinara, E., and Mcdonnell, R. (2015). "Animation realism affects perceived character appeal of a self-virtual face," in *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games* (New York, NY: Association for Computing Machinery).

Kopp, S., Krenn, B., Marsella, S., Marshall, A. N., Pelachaud, C., Pirker, H., et al. (2006). "Towards a common framework for multimodal generation: the behavior markup language," in *Intelligent Virtual Agents*, eds. J. Gratch, M. Young, R. Aylett, D. Ballin, P. and Olivier (Berlin: Springer Berlin Heidelberg), 205–217.

Kruzic, C. O., Kruzic, D., Herrera, F., and Bailenson, J. N. (2020). Facial expressions contribute more than body movements to conversational outcomes in avatar-mediated virtual environments. *Sci. Rep.* 10:4. doi: 10.1038/s41598-020-76672-4

Kucherenko, T., Nagy, R., Yoon, Y., Woo, J., Nikolov, T., Tsakov, M., et al. (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, ICMI '23* (New York, NY: Association for Computing Machinery), 792–801.

Lea, C., Flynn, M. D., Vidal, R., Reiter, A., and Hager, G. D. (2017). "Temporal convolutional networks for action segmentation and detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI: IEEE).

Li, J., Kang, D., Pei, W., Zhe, X., Zhang, Y., Bao, L., et al. (2023). *Audio2gestures: Generating Diverse Gestures from Audio* (Los Alamitos, CA: IEEE Institute of Electrical and Electronics Engineers).

Li, R., Yang, S., Ross, D. A., and Kanazawa, A. (2021). *Ai Choreographer: Music Conditioned 3D Dance Generation With AIST++.*

Li, T., Bolkart, T., Black, M. J., Li, H., and Romero, J. (2017). Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* 36, 194:1–194:17. doi: 10.1145/3130800.3130813

Liu, H. (2024). *Emotion Detection Through Body Gesture and Face.*

Liu, H., Iwamoto, N., Zhu, Z., Li, Z., Zhou, Y., Bozkurt, E., et al. (2022a). "Disco: Disentangled implicit content and rhythm learning for diverse co-speech gestures synthesis," in *Proceedings of the 30th ACM International Conference on Multimedia* (New York: ACM) 3764–3773.

Liu, H., Zhu, Z., Becherini, G., Peng, Y., Su, M., Zhou, Y., et al. (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 (CVPR)* (Seattle, WA: IEEE), 1144–1154.

Liu, H., Zhu, Z., Iwamoto, N., Peng, Y., Li, Z., Zhou, Y., et al. (2022b). "Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis," in *European Conference on Computer Vision* (Cham: Springer).

Lombard, M., Ditton, T. B., and Weinstein, L. (2009). *Measuring Presence: The Temple Presence Inventory.* Available online at: https://matthewlombard.com/research/p2_ab.html (Accessed July 21, 2025).

Ma, F., Xie, Y., Li, Y., He, Y., Zhang, Y., Ren, H., et al. (2025). A review of human emotion synthesis based on generative technology. *IEEE Trans. Affect. Comput.* 2025, 1–20. doi: 10.1109/TAFFC.2025.3573878

Maiorca, A., Yoon, Y., and Dutoit, T. (2022). *Evaluating the Quality of a Synthesized Motion with the Frchet Motion Distance* (New York, NY: Association for Computing Machinery).

Mal, D., Wolf, E., Döllinger, N., Botsch, M., Wienrich, C., and Latoschik, M. E. (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, CHI EA '24* (New York, NY: Association for Computing Machinery).

Marinetti, C., Moore, P., Lucas, P., and Parkinson, B. (2011). *Emotions in Social Interactions: Unfolding Emotional Experience*, 31–46.

Marsella, S., Xu, Y., Lhommet, M., Feng, A., Scherer, S., and Shapiro, A. (2013). "Virtual character performance from speech," in *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA '13* (New York, NY: Association for Computing Machinery), 25–35.

McNeill, D. (1992). "Hand and mind1," in *Advances in Visual Semiotics*, 351.

Microsoft Mesh (2025). *Microsoft Mesh.* Available online at: https://www.microsoft.com/en-us/microsoft-teams/microsoft-mesh (Accessed July 19, 2024).

Mughal, M. H., Dabral, R., Habibie, I., Donatelli, L., Habermann, M., and Theobalt, C. (2024). "Convofusion: Multi-modal conversational diffusion for co-speech gesture synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA: IEEE), 1388–1398.

NVIDIA ACE (2025). *NVIDIA ACE.* Available onlne at: https://developer.nvidia.com/ace (Accessed July 19, 2024).

Osimo, S. A., Pizarro, R., Spanlang, B., and Slater, M. (2015). Conversations between self and self as sigmund freuda virtual body ownership paradigm for self counselling. *Sci. Rep.* 5:srep13899. doi: 10.1038/srep13899

Özyürek, A. (2014). Hearing and seeing meaning in speech and gesture: Insights from brain and behavior. *Philosoph. Trans. Royal Soc. B: Biol. Sci.* 369:20130296. doi: 10.1098/rstb.2013.0296

Pan, X., and Hamilton, A. (2018). Why and how to use virtual reality to study human social interaction: the challenges of exploring a new research landscape. *Br. J. Psychol.* 109, 395–417. doi: 10.1111/bjop.12290

Patotskaya, Y., Hoyet, L., Olivier, A.-H., Pettré, J., and Zibrek, K. (2023). Avoiding virtual humans in a constrained environment: Exploration of novel behavioral measures. *Comput. Graph.* 110, 162–172. doi: 10.1016/j.cag.2023.01.001

Patterson, M. L., Fridlund, A. J., and Crivelli, C. (2023). Four misconceptions about nonverbal communication. *Persp. Psychol. Sci.* 18, 1388–1411. doi: 10.1177/17456916221148142

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A. A., Tzionas, D., et al. (2019). "Expressive body capture: 3D hands, face, and body from a single image," in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA: IEEE Institute of Electrical and Electronics Engineers), 10975–10985.

Petrovich, M., Black, M. J., and Varol, G. (2021). "Action-conditioned 3D human motion synthesis with transformer VAE," in *International Conference on Computer Vision (ICCV)* (Montreal, QC: IEEE Institute of Electrical and Electronics Engineers).

Pham, H. X., Cheung, S., and Pavlovic, V. (2017a). "Speech-driven 3D facial animation with implicit emotional awareness: a deep learning approach," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017* (Honolulu, HI: IEEE Computer Society), 2328–2336.

Pham, H. X., Wang, Y., and Pavlovic, V. (2017b). End-to-end learning for 3d facial animation from raw waveforms of speech. *arXiv*. doi: 10.1145/3242969.3243017

PlayAI (2025). *AI Voice Generator & Text to Speech AI Voice Platform*. Available online at: https://play.ht/ (Accessed July 21, 2025).

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *Am. Scient.* 89, 344–350. doi: 10.1511/2001.28.344

Poggi, I., Pelachaud, C., de Rosis, F., Carofiglio, V., and De Carolis, B. (2005). *Greta. A Believable Embodied Conversational Agent* (Dordrecht: Springer Netherlands).

Qi, X., Liu, C., Sun, M., Li, L., Fan, C., and Yu, X. (2023). *Diverse 3D Hand Gesture Prediction from Body Dynamics by Bilateral Hand Disentanglement* (Vancouver, BC: IEEE Institute of Electrical and Electronics Engineers).

Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.-Y., Johnson, J., et al. (2020). Accelerating 3D deep learning with pytorch3d. *arXiv* [preprint] arXiv:2007.08501. doi: 10.48550/arXiv.2007.08501

Reeves, B., and Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and PLA*. Chicago, IL: Bibliovault OAI Repository, the University of Chicago Press.

Replika (2025). *The AI Companion Who Cares*. Available online at: https://replika.com/ (Accessed July 19, 2024).

Richard, A., Zollhöfer, M., Wen, Y., de la Torre, F., and Sheikh, Y. (2021). "Meshtalk: 3D face animation from speech using cross-modality disentanglement," in Proceedings *of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal, QC: IEEE), 1173–1182. doi: 10.1109/ICCV48922.2021.00121

Rizzolatti, G., and Arbib, M. A. (1998). Language within our grasp. *Trends Neurosci.* 21, 188–194. doi: 10.1016/S0166-2236(98)01260-0

Rogers, S. L., Broadbent, R., Brown, J., Fraser, A., and Speelman, C. P. (2021). Realistic motion avatars are the future for social interaction in virtual reality. *Front. Virtual Reality.* 2:750729. doi: 10.3389/frvir.2021.750729

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA: IEEE), 10684–10695. doi: 10.1109/CVPR52688.2022.01042

Rosenthal-von der Pütten, A. M., Krmer, N., Gratch, J., and Kang, S.-H. (2010). "It doesn't matter what you are!" explaining social effects of agents and avatars. *Comput. Human Behav.* 26, 1641–1650. doi: 10.1016/j.chb.2010.06.012

Roth, D., Kleinbeck, C., Feigl, T., Mutschler, C., and Latoschik, M. E. (2018a). "Beyond replication: augmenting social behaviors in multi-user virtual realities," in *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (Tuebingen/Reutlingen: IEEE), 215–222. doi: 10.1109/VR.2018.8447550

Roth, D., Kullmann, P., Bente, G., Gall, D., and Latoschik, M. E. (2018b). "Effects of hybrid and synthetic social gaze in avatar-mediated interactions," in *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)* (Munich: IEEE), 103–108. doi: 10.1109/ISMAR-Adjunct.2018.00044

Russell, J. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.* 39, 1161–1178. doi: 10.1037/h0077714

Schlosberg, H. (1954). Three dimensions of emotion. *Psychol. Rev.* 61, 81–88. doi: 10.1037/h0054570

Sharkov, F., Silkin, V., and Kireeva, O. (2022). Non-verbal signs of personality: Communicative meanings of facial expressions. *RUDN J. Sociol.* 22, 387–403. doi: 10.22363/2313-2272-2022-22-2-387-403

Smith, H. J., and Neff, M. (2018). "Communication behavior in embodied virtual reality," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18* (New York, NY: Association for Computing Machinery).

Stewart, C., Mitchell, D., Pasternak, S., Tremblay, P., and Finger, E. (2024). The nonverbal expression of guilt in healthy adults. *Sci. Rep.* 14:980. doi: 10.1038/s41598-024-60980-0

Synthesia (2025). *Turn Text to Video, in Minutes*. Available online at: https://www.synthesia.io/ (Accessed July 19, 2024).

Tan, S., and Nareyek, A. (2009). *Integrating Facial, Gesture, and Posture Emotion Expression for a 3D Virtual Agent* (Wolverhampton: The University of Wolverhampton).

Taylor, S. L., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A. G., et al. (2017). A deep learning approach for generalized speech animation. *ACM Trans. Graph.* 36, 93:1–93:11. doi: 10.1145/3072959.3073699

Thiébaux, M., Marsella, S., Marshall, A. N., and Kallmann, M. (2008). "Smartbody: behavior realization for embodied conversational agents," in *Adaptive Agents and Multi-Agent Systems* (Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems).

Thomas, S., Ferstl, Y., McDonnell, R., and Ennis, C. (2022). "Investigating how speech and animation realism influence the perceived personality of virtual characters and agents," in *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (Christchurch: IEEE), 11–20.

Tim Sweeney (2025). *Unreal Engine 5.4 Documentation*. Available online at: https://www.unrealengine.com/en-US (Accessed July 19, 2024).

Ton Roosendaal (2025). *Blender Foundation Documentation*. Available online at: https://docs.blender.org/ (Accessed July 19, 2024).

Tracy, J. L., Randles, D., and Steckler, C. M. (2015). The nonverbal communication of emotions. *Curr. Opini. Behav. Sci. Soc. Behav.* 3, 25–30. doi: 10.1016/j.cobeha.2015.01.001

Unity Technologies (2025). *Documentation*. Available online at: https://docs.unity.com/ (Accessed July 19, 2024).

Valle-Pérez, G., Henter, G. E., Beskow, J., Holzapfel, A., Oudeyer, P.-Y., and Alexanderson, S. (2021). Transflower. *ACM Trans. Graph.* 40, 1–14. doi: 10.1145/3478513.3480570

Van Den Oord, A., Vinyals, O., and Kavukcuoglu, K. (2017). Neural discrete representation learning. *Adv. Neural Inf. Process. Syst*, 30, 6309–6318. Available online at: https://dl.acm.org/doi/10.5555/3295222.3295378

Vasiliu, M. M., Guarese, R., Jaatinen, J., Johnson, F., Edvinsson, B., and Romero, M. (2025). "Towards enhancing industrial training through conversational AI," in CUI '24: Proceedings of the 6th ACM Conference on Conversational User Interfaces (New York: ACM).

Vaswani, A. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems* (Red Hook, NY: Curran Associates Inc.).

Watson, D. B. and Tellegen, A. (1985). Toward a consensual structure of mood. *Psychol. Bullet.* 982, 219–35.

Wobbrock, J. O., Findlater, L., Gergle, D., and Higgins, J. J. (2011). *The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures.* (New York, NY: Association for Computing Machinery), 143–146.

Xing, J., Xia, M., Zhang, Y., Cun, X., Wang, J., and Wong, T.-T. (2023). Codetalker: Speech-driven 3d facial animation with discrete motion prior. *arXiv* [preprint] arXiv:2301.02379. doi: 10.1109/CVPR52729.2023.01229

Yang, S., Wu, Z., Li, M., Zhang, Z., Hao, L., Bao, W., et al. (2023a). "Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models," in *Proceedings of the 32nd International Joint Conference on Artificial Intelligence, IJCAI 2023*.

Yang, S., Wu, Z., Li, M., Zhang, Z., Hao, L., Bao, W., et al. (2023b). "Qpgesture: Quantization-based and phase-guided motion matching for natural speech-driven gesture generation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR* (Los Alamitos, CA: IEEE Computer Society), 2321–2330.

Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., et al. (2023). "Generating holistic 3D human motion from speech," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Amsterdam: SSRN), 469–480.

Yin, L., Wang, Y., He, T., Liu, J., Zhao, W., Li, B., et al. (2023). *Emog: Synthesizing Emotive Co-Speech 3D Gesture With Diffusion Model*. Amsterdam.

Yin, T., Hoyet, L., Christie, M., Cani, M.-P., and Pettré, J. (2022). The one-man-crowd: Single user generation of crowd motions using virtual reality. *IEEE Trans. Vis. Comput. Graph.* 28, 2245–2255. doi: 10.1109/TVCG.2022.3150507

Yin, T., Hoyet, L., Christie, M., Cani, M.-P., and Pettré, J. (2024). With or without you: Effect of contextual and responsive crowds on vr-based crowd motion capture. *IEEE Trans. Vis. Comput. Graph.* 30, 2785–2795. doi: 10.1109/TVCG.2024.3372038

Yoon, Y., Cha, B., Lee, J.-H., Jang, M., Lee, J., Kim, J., et al. (2020). Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Trans. Graph.* 39:6. doi: 10.1145/3414685.3417838

Zhang, J., Brandsttter, K., and Steed, A. (2023). "Supporting co-presence in populated virtual environments by actor takeover of animated characters," in *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (Sydney: IEEE), 940–949. doi: 10.1109/ISMAR59233.2023.00110

Zhao, W., Riggs, K., Schindler, I., and Holle, H. (2018). Transcranial magnetic stimulation over left inferior frontal and posterior temporal cortex disrupts gesture-speech integration. *J. Neurosci.* 38, 1891–1900. doi: 10.1523/JNEUROSCI.1748-17.2017

Zhou, Y., Xu, Z., Landreth, C., Kalogerakis, E., Maji, S., and Singh, K. (2018). Visemenet: Audio-driven animator-centric speech animation. *ACM Trans. Graph.* 37:4. doi: 10.1145/3197517.3201292

Zhu, X., Gong, Y., Xu, T., Lian, W., Xu, S., and Fan, L. (2023). Incongruent gestures slow the processing of facial expressions in university students with social anxiety. *Front. Psychol.* 14, 1–9. doi: 10.3389/fpsyg.2023.1199537

Zibrek, K., Kokkinara, E., and Mcdonnell, R. (2018). The effect of realistic appearance of virtual characters in immersive environments - does the character's personality play a role? *IEEE Trans. Visualizat. Comp. Graph.* 24, 1681–1690. doi: 10.1109/TVCG.2018.2794638