Human-centered Evaluation of Generative Models for Emotional 3D Animation Generation in VR

Anonymous ICCV submission

Paper ID 17

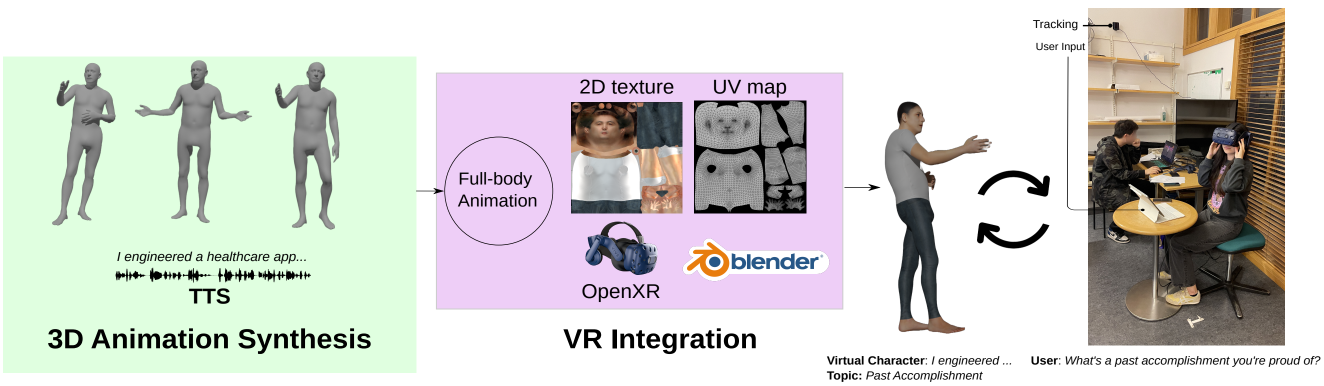

Figure 1. **Human-centered Evaluation for Emotional 3D Animation in VR.** Participants interact with a virtual character using a VR headset in a modular setup that supports various text-to-speech (TTS) and speech-driven 3D animation methods. The system generates 3D facial and body animations from TTS speech segments, maps them onto a textured character via UV mapping, and renders them in real-time using Blender (OpenXR). Participants' positions are tracked via base stations, and a tablet is used for in-session feedback.

Abstract

Facial expressions and body gestures are vital for conveying emotion in social interaction. While generative models can produce speech-synchronized 3D animations, traditional 2D evaluations often miss user-perceived emotional quality. We present a VR-based user study ($N = 48$) evaluating three state-of-the-art speech-driven 3D animation models across two emotions—happiness (high arousal) and neutral (mid arousal)—using user-centric metrics: arousal realism, naturalness, enjoyment, diversity, and interaction quality. We also compare against real human expressions generated via a reconstruction-based method. Models explicitly encoding emotion achieved higher recognition rates than those driven solely by speech. Happy animations were rated significantly more realistic and natural than neutral ones, highlighting challenges in modeling subtle emotion. Generative models underperformed compared to reconstructions in facial expression quality, and all received comparable ratings for enjoyment and interaction quality. Users reliably recognized gesture diversity across generative outputs, motivating integration of perceptual feedback into animation models.

1. Introduction

Realistic VR interactions depend on generating expressive verbal and non-verbal behaviors, such as gestures and facial expressions [19, 24, 35]. These cues are vital for conveying emotion [7, 34], yet challenging to synthesize convincingly. Early systems used rule-based or motion-capture approaches [3, 17], but recent generative models enable scal-

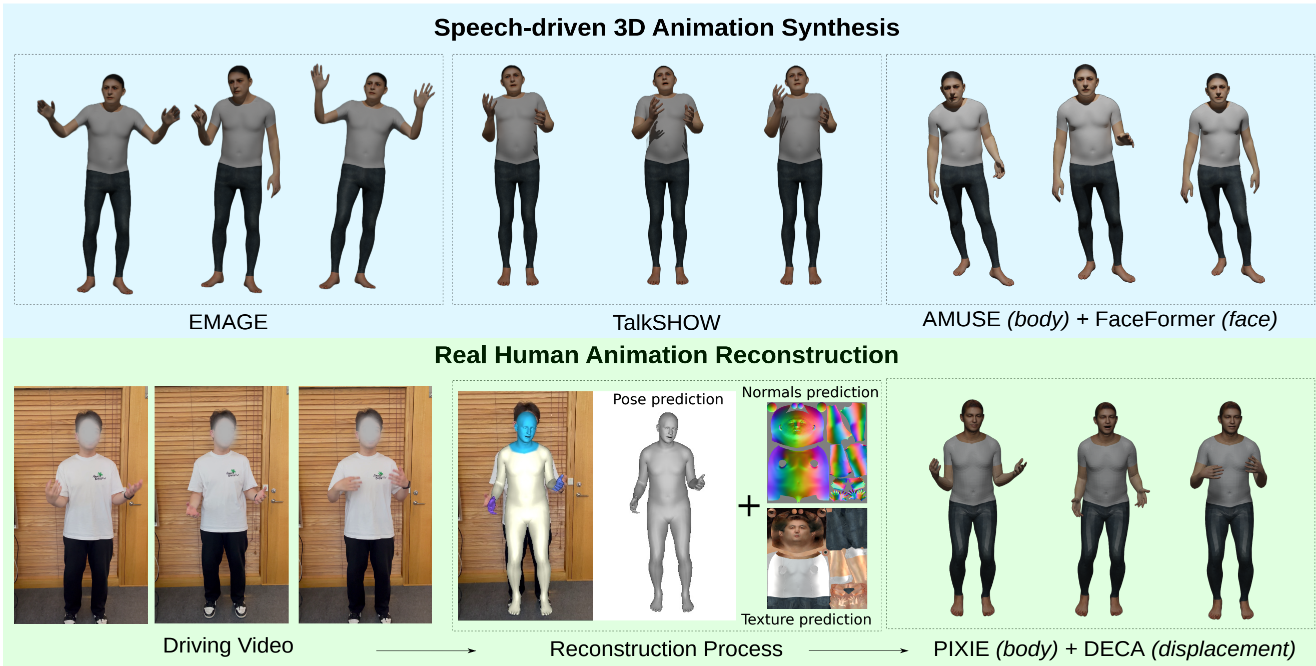

Figure 2. **Qualitative comparison of generative and real animations.** Top: Sample frames from three generative models—EMAGE [26], TalkSHOW [38], and AMUSE (body) + FaceFormer (face) [4, 12]. Bottom: Reconstruction-based baseline using video input. PIXIE [14] and DECA [15] extract pose, expression, and texture, which are rendered per frame to create high-fidelity human-like animations.

able, speech-driven 3D animation [4, 6, 8, 38]. However, most evaluations rely on objective metrics [25, 40], overlooking user perception. Studies rarely assess full-body, emotionally rich animations in real-time VR dialogue [5, 9, 10].

We address this gap via a VR-based user study ($N = 48$) comparing three speech-driven generative models—AMUSE [4], TalkSHOW [38], and EMAGE [26]—and FaceFormer [12] for facial animation, and PIXIE [14] as a real-human baseline. Using SMPL-X avatars [28], we evaluate two arousal levels (happy, neutral) across five perceptual metrics: realism, naturalness, enjoyment, diversity, and interaction quality. Our contributions are: (1) A perceptual evaluation of emotional 3D animation in immersive, real-time VR dialogue; (2) Comparative user study of generative vs. real-human animation; (3) Analysis of strengths and limitations in current models for expressive interaction.

2. Related Work

Social Interaction in VR. Human communication relies on tightly coupled speech and gestures, which share sensorimotor representations [1, 18] and can complement or replace each other [22, 27]. Emotion modeling often uses categorical (e.g., Ekman [11]) or dimensional (e.g., arousal–valence [33]) frameworks. We adopt the dimension framework and validate perception via an Ekman-style classifier. In VR, character animation has traditionally relied on rule-based or teleoperated systems [3, 17], while emerging platforms (e.g., Synthesia, Replika) offer expressiveness with limited user controllability. Previous work has explored rendering, animation, and social cues [5, 21, 23, 31], although often without real-time generative control.

Generative Animation Models. Recent models synthesize speech-driven 3D facial and body animation [13, 20, 37], with works using SMPL-X [28] meshes. Emotion-aware animation generation is evolving [4, 39], but few studies evaluate these in real-time VR. Our work combines emotional TTS and generative animation in VR, enabling perceptual evaluation of interaction quality.

3. Implementation Details

System Overview. We implement a modular VR pipeline integrating speech-driven 3D animation models with TTS and real-time rendering in Blender. Generated animations are mapped onto a textured SMPL-X [28] avatar and streamed to participants using an HTC Vive Pro 2 via Blender's OpenXR interface.

Generative Models. We evaluate three state-of-the-art models: EMAGE [26], TalkSHOW [38], and AMUSE [4]. AMUSE is combined with FaceFormer [12] to enable full-body animation. All models generate 3D facial and body motion from speech. EMAGE uses a rhythm-aware TCN and VQ-VAE, TalkSHOW applies Wav2Vec features with a split architecture (VQ-VAE for body, transformer for face), while AMUSE uses a ViT-based feature encoder and conditional diffusion model for emotion-aware gesture generation. FaceFormer provides frame-level facial expressions via autoregressive transformers.

Animation Integration. Model outputs (pose and expression parameters) are retargeted to a SMPL-X avatar with consistent shape and texture across methods. FaceFormer outputs are converted from FLAME topology to SMPL-X expression parameters via optimization and then aligned frame-wise with body gestures. For all methods, dialogue responses are templated, converted to speech using PlayHT TTS [29], and used as driving audio input.

Real Human Baseline. We compare generative animations against a reconstruction-based method using real human video input. We employ PIXIE [14] for body and facial parameter estimation and DECA [16] for high-fidelity facial displacement. Sequences are rendered using per-frame UV-mapped textures and lighting, exported via PyTorch3D [30], and animated in Blender using geometry nodes (Fig. 2).

Rendering Setup. All animations are rendered with consistent camera, lighting, and background using Blender 3.4. The SMPL-X add-on handles mesh import, rigging, and real-time playback. Audio were sampled at 16 kHz, and models were run with default hyperparameters.

4. Human-centered Evaluation

4.1. Research Questions

We examine six research questions comparing happy and neutral animations: (RQ1) Which method yields the highest perceived realism during interaction? (RQ2) Which model produces the most natural facial expressions and body gestures? (RQ3) Do methods affect perceived enjoyment? (RQ4) Do they differ in interaction quality? (RQ5) Can users perceive motion diversity when shown two neutral animations of the same utterance? (RQ6) Can participants correctly identify the intended arousal level of a given animation?

4.2. User Study

Participants and Setup. We recruited 48 participants (28M, 20F; age 19–48, M = 26.7, SD = 5.3) via university channels. Most (70.8%) had played video games in the past year, and their prior VR experience ranged from below average (6.3%) to very good (22.9%). All gave informed consent and received gift cards. The study was approved by the local ethics board. Participants wore an HTC Vive Pro 2 headset (2448×2448 per eye, 90Hz), tracked via SteamVR base stations. The VR environment was rendered in Blender 3.4 with OpenXR, running on a workstation (i9-13900K, 64GB RAM, RTX A6000). Animations were pre-generated to ensure synchronized playback during interaction.

Design and Procedure. We employed a within-subject design with two factors: *method* (EMAGE, TalkSHOW, AMUSE+FaceFormer, PIXIE+DECA) and *scenario* (Happy, Neutral, Diversity), totaling 12 conditions per participant. HEA and NEA involved short conversations reflecting high or mid arousal; DV showed two agents performing the same utterance with varied gestures. Prompts/responses were scripted and counterbalanced using a Latin Square. Participants read a prompt, wore the headset to view the animation, then removed it to complete a brief survey. This was repeated for all 12 trials, with additional pre- and post-study questionnaires on demographics and overall experience.

Measures. A 21-item questionnaire assessed realism, facial/body naturalness, interaction quality, emotion recognition, diversity, and social presence using Likert scales (5-point, 3-point for arousal, binary for diversity). Items were adapted from prior VR and animation studies [2, 17, 32].

Analysis. Due to non-normality, we used Aligned Rank Transform (ART) ANOVA [36], with Bonferroni corrections for pairwise comparisons.

5. Results

Perceived Realism, Naturalness, and Enjoyment. Animations expressing happy emotion were rated significantly

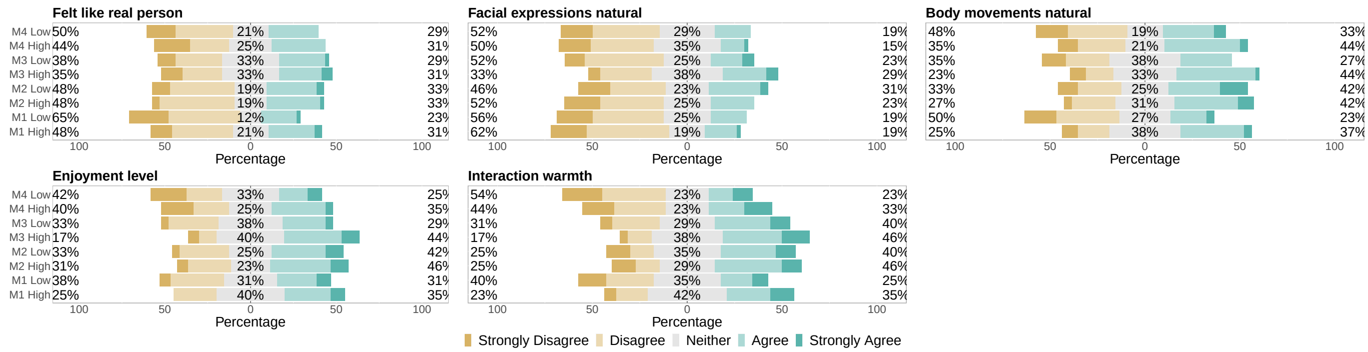

Figure 3. **Summary of Likert ratings.** Ratings for Animation Realism, Naturalness (face/body), Enjoyment, and Interaction Quality. EMAGE, TalkSHOW, PIXIE+DECA, and AMUSE+FaceFormer are denoted as M1–M4; 'High' and 'Low' indicate happy and neutral.

more realistic and natural than neutral ones ($p < 0.001$ and $p = 0.01$, respectively), with no overall method effect for realism or body gestures. For facial expressions, PIXIE+DECA outperformed EMAGE ($p = 0.01$), especially in neutral conditions. EMAGE received lower facial ratings for neutral, while TalkSHOW performed better in the happy condition. Enjoyment ratings showed no significant differences across methods or arousal levels and were generally neutral.

Interaction Quality and Emotional Recognition. TalkSHOW was rated highest in interaction quality, significantly outperforming AMUSE+FaceFormer ($p = 0.027$), with no other significant method or emotion effects. Emotion recognition accuracy was higher for neutral (79%) than for happy (61%). AMUSE+FaceFormer had the highest high-arousal recognition, while PIXIE+DECA led for neutral, suggesting real human reconstructions better convey subtle emotions, while emotion-aware generative models better support high-arousal expression.

Motion Diversity and Overall Impressions. AMUSE+FaceFormer showed the highest perceived diversity (96%), with EMAGE lowest (71%). TalkSHOW and PIXIE+DECA fell in between (79%), aligning with computed joint-space diversity metrics (2-norm: AMUSE+FaceFormer 2.94, EMAGE 2.53, TalkSHOW 2.08). PIXIE+DECA showed no diversity due to its deterministic reconstruction. Post-study ratings indicated PIXIE+DECA was most favored for realism and facial quality, while AMUSE+FaceFormer maintained balanced impressions. EMAGE and TalkSHOW were perceived as lower in social closeness, highlighting the advantage of reconstruction-based methods in conveying subtle emotion and presence. All user ratings are summarized in Fig. 3.

6. Discussion and Conclusion

Our study reveals that perceived animation quality varies significantly with emotional arousal. High-arousal (happy) animations were rated as more realistic and natural than neutral ones, with AMUSE+FaceFormer and PIXIE+DECA leading in emotion recognition accuracy. PIXIE+DECA produced the most natural facial expressions—particularly for subtle emotions—but its reliance on real video input and long inference time (412s for 10s generation) limits scalability. AMUSE+FaceFormer achieved strong arousal recognition and high diversity, balancing expressiveness with a moderate runtime (8.5s). TalkSHOW (20.3s), though lower in emotional expressiveness, ranked highest in interaction quality. EMAGE (0.8s), while the fastest, was the least diverse. Participants identified neutral arousal more accurately overall, with mid-arousal gestures proving easier to interpret across models.

Across methods, animation diversity was best perceived in AMUSE+FaceFormer (96%) and lowest in EMAGE (71%), aligning with quantitative diversity scores. All generative models showed potential for creating believable agents, though enjoyment and interaction quality remained limited compared to human-based animation. These findings highlight the importance of combining perceptual user studies with technical evaluation to guide the development of expressive, emotionally intelligent virtual characters for immersive social interaction.

References

[1] Michael Andric, Ana Solodkin, Giovanni Buccino, Susan Goldin-Meadow, Giacomo Rizzolatti, and Steven L Small. Brain function overlaps when people observe emblems, speech, and grasping. *Neuropsychologia*, 51(8):1619–1629, 2013. 2

[2] Frank Biocca, Chad Harms, and Judee K. Burgoon. Toward a More Robust Theory and Measure of Social Presence: Review and Suggested Criteria. *Presence: Teleoperators and Virtual Environments*, 12(5):456–480, 2003. 3

[3] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy W. Bickmore. Beat: the behavior expression animation toolkit. *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001. 1, 2

[4] Kiran Chhatre, Radek Daněček, Nikos Athanasiou, Giorgio Becherini, Christopher Peters, Michael J. Black, and Timo Bolkart. 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)*, pages 1942–1953, 2024. 2, 3

[5] Kiran Chhatre, Renan Guarese, Andrii Matviienko, and Christopher Peters. 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, 2025. Association for Computing Machinery. 2

[6] Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. Emotional speech-driven animation with content-emotion disentanglement. ACM, 2023. 2

[7] Elisa De Stefani and Doriana De Marco. Language, gesture, and emotional communication: An embodied view of social interaction. *Frontiers in Psychology*, 10, 2019. 1

[8] Anna Deichler, Kiran Chhatre, Christopher Peters, and Jonas Beskow. Spatio-temporal priors in 3d human motion. 2021. 2

[9] Anna Flavia Di Natale, Matilde Ellen Simonetti, Stefania La Rocca, and Emanuela Bricolo. Uncanny valley effect: A qualitative synthesis of empirical research to assess the suitability of using virtual faces in psychological research. *Computers in Human Behavior Reports*, 10:100288, 2023. 2

[10] Haoyang Du, Kiran Chhatre, Christopher Peters, Brian Keegan, Rachel McDonnell, and Cathy Ennis. Synthetically expressive: Evaluating gesture and voice for emotion and empathy in vr and 2d scenarios, 2025. 2

[11] Paul Ekman. Facial expression and emotion. *American Psychologist*, 48(4):384–392, 1993. 2

[12] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. *arXiv preprint arXiv:2112.05329*, 2021. 2, 3

[13] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18749–18758. IEEE, 2022. 2

[14] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, 2021. 2, 3

[15] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. 2021. 2

[16] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (ToG), Proc. SIGGRAPH*, 40(4):88:1–88:13, 2021. 3

[17] Alan Fraser, Isabella Branson, Ross Hollett, Craig Speelman, and Shane Rogers. Expressiveness of real-time motion captured avatars influences perceived animation realism and perceived quality of social interaction in virtual reality. *Frontiers in Virtual Reality*, 3:981400, 2022. 1, 2, 3

[18] Maurizio Gentilucci, Paolo Bernardis, Girolamo Crisi, and Riccardo Dalla Volta. Repetitive transcranial magnetic stimulation of broca's area affects verbal responses to gesture observation. *Journal of Cognitive Neuroscience*, 18(7):1059–1074, 2006. 2

[19] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F. Troje, and Marc-André Carbonneau. Zeroeggs: Zero-shot example-based gesture generation from speech. *Computer Graphics Forum*, 42(1):206–216, 2023. 1

[20] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019. 2

[21] Rosanna E. Guadagno, Jim Blascovich, Jeremy N. Bailenson, and Cade Andrew McCall. Virtual humans and persuasion: The effects of agency and behavioral realism. *Media Psychology*, 10:1 – 22, 2007. 2

[22] Thomas C Gunter and Patric Bach. Communicating hands:

Erps elicited by meaningful symbolic hand postures. *Neuroscience letters*, 372(1-2):52–56, 2004. 2

[23] Elena Kokkinara and Rachel Mcdonnell. Animation realism affects perceived character appeal of a self-virtual face. *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games*, 2015. 2

[24] Catherine Oh Kruzic, David Kruzic, Fernanda Herrera, and Jeremy N. Bailenson. Facial expressions contribute more than body movements to conversational outcomes in avatar-mediated virtual environments. *Scientific Reports*, 10, 2020. 1

[25] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++, 2021. 2

[26] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J. Black. 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)*, pages 1144–1154, 2024. 2, 3

[27] Aslı Özyürek. Hearing and seeing meaning in speech and gesture: Insights from brain and behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369 (1651):20130296, 2014. 2

[28] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2, 3

[29] PlayHT. AI Voice Generator: Realistic text to speech v2.0, 2025. 3

[30] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*, 2020. 3

[31] Byron Reeves and Clifford Nass. The media equation: How people treat computers, television, and new media like real people and pla. *Bibliovault OAI Repository, the University of Chicago Press*, 1996. 2

[32] Shane L. Rogers, Rebecca Broadbent, Jemma Brown, Alan Fraser, and Craig P. Speelman. Realistic motion avatars are the future for social interaction in virtual reality. In *Frontiers in Virtual Reality*, 2021. 3

[33] James Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 1980. 2

[34] Jessica L Tracy, Daniel Randles, and Conor M Steckler. The nonverbal communication of emotions. *Current Opinion in Behavioral Sciences*, 3:25–30, 2015. Social behavior. 1

[35] Miruna Maria Vasiliu, Renan Guarese, Jonas Jaatinen, Fabian Johnson, Benjamin Edvinsson, and Mario Romero. Towards enhancing industrial training through conversational AI. In *CUI '24: Proceedings of the 6th ACM Conference on Conversational User Interfaces*, 2025. 1

[36] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. *The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures*, page 143–146. Association for Computing Machinery, New York, NY, USA, 2011. 3

[37] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, and Haolin Zhuang. 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*, pages 2321–2330. IEEE, 2023. 2

[38] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 469–480, 2023. 2, 3

[39] Lianying Yin, Yijun Wang, Tianyu He, Jinming Liu, Wei Zhao, Bohan Li, Xin Jin, and Jianxin Lin. Emog: Synthesizing emotive co-speech 3d gesture with diffusion model, 2023. 2

[40] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics*, 39(6), 2020. 2