Evaluating Speech and Video Models for Face-Body Congruence

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

Renan Guarese
KTH Royal Institute of Technolgy
Stockholm, Sweden
guarese@kth.se

Andrii Matviienko
KTH Royal Institute of Technolgy
Stockholm, Sweden
andriim@kth.se

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

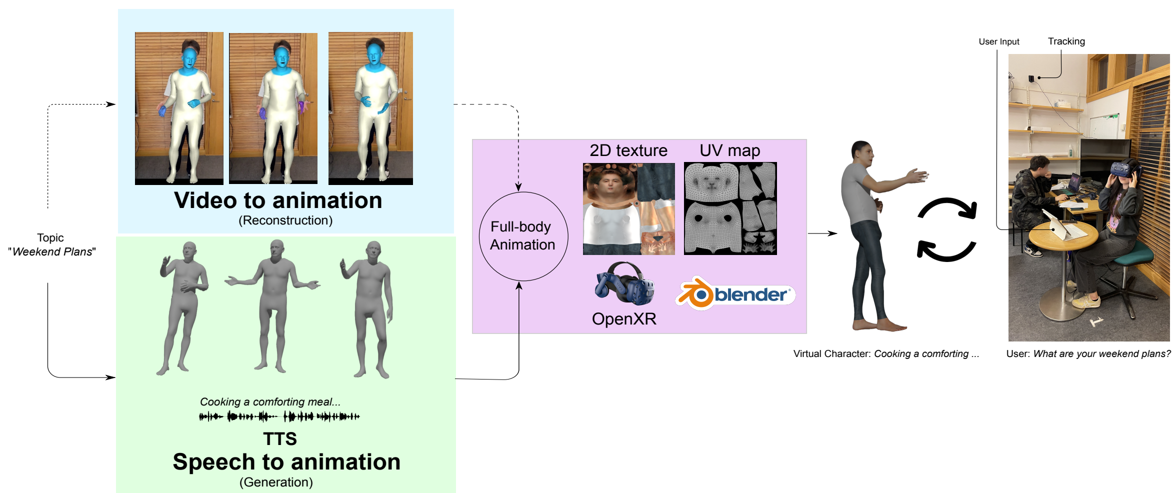

Figure 1: Left: (1) Speech-driven method (AMUSE + FaceFormer) generates speech-synchronized full-body animation; (2) Video-driven reconstruction (PIXIE + DECA) tracks face, hands, and body. Center: Motion data is mapped onto a textured 3D character, rendered, and streamed via Blender for VR interaction. Right: Participant interacting with virtual character.

Abstract

Animations produced by generative models are often evaluated using objective quantitative metrics that do not fully capture perceptual effects in immersive virtual environments. To address this gap, we present a preliminary perceptual evaluation of generative models for animation synthesis, conducted via a VR-based user study (N = 48). Our investigation specifically focuses on animation congruency—ensuring that generated facial expressions and body gestures are both congruent with and synchronized to driving speech. We evaluated two state-of-the-art methods: a speech-driven full-body animation model and a video-driven full-body reconstruction model, assessing their capability to produce congruent facial expressions and body gestures. Our results demonstrate a strong user preference for combined facial and body animations, highlighting that congruent multimodal animations significantly enhance perceived realism compared to animations featuring only a single modality. By incorporating VR-based perceptual feedback into training pipelines, our approach provides a foundation for developing more engaging and responsive virtual characters.

I3D Companion '25, Jersey City, NJ, USA

ACM ISBN /25/05
https://doi.org/10.1145/3722564.3728374

CCS Concepts

• **Computing methodologies** → **Animation**.

Keywords

Generative Models, Conversational Agents, Virtual Reality

ACM Reference Format:
Kiran Chhatre, Renan Guarese, Andrii Matviienko, and Christopher Peters. 2025. Evaluating Speech and Video Models for Face-Body Congruence. In *Companion Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games (I3D Companion '25), May 07–09, 2025, Jersey City, NJ, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3722564.3728374

1 Introduction

Immersive VR experiences rely on natural conversational interactions between users and virtual characters, requiring accurate replication of speech, gestures, and facial expressions. Non-verbal cues such as gesture emotion, synchrony between facial expressions and body gestures, and eye contact are essential for conveying emotions and guiding responses [Sharkov et al. 2022; Stewart et al. 2024]. However, ensuring that generated animations capture these details remains challenging [Patterson et al. 2023].

Existing generative methods have typically addressed facial animation [Cudeiro et al. 2019; Fan et al. 2022; Pham et al. 2017; Richard et al. 2021; Xing et al. 2023] and body gestures [Ginosar et al. 2019; Habibie et al. 2022; Liang et al. 2022; Yang et al. 2023;

Yoon et al. 2020] separately. Only recently have methods jointly tackled facial and gesture synthesis [Habibie et al. 2022; Liu et al. 2024; Mughal et al. 2024; Ng et al. 2024; Shi et al. 2024; Yi et al. 2023]. Despite these developments, to our knowledge, no prior studies have evaluated the perceived congruency—defined as the synchrony between facial animations and body gestures. Existing evaluations of full-body animations [Alexanderson et al. 2023; Chhatre et al. 2024; Daněček et al. 2023] mainly rely on objective metrics like Fréchet Motion Distance [Yoon et al. 2020], beat alignment [Li et al. 2021], and gesture diversity [Li et al. 2023], overlooking subjective aspects like perceived face-body congruency. Although previous studies, such as the GENEA Challenge [Kucherenko et al. 2023] and AV-Flow [Chatziagapi et al. 2025], have assessed virtual faces and gestures separately, a comprehensive 3D evaluation is lacking. Deichler et al. [Deichler et al. 2024] evaluated animations in 2D and VR from a third-person perspective but did not address face-body congruency. We address this gap through a VR-based perceptual study (N=48), examining the congruency between facial expressions and body gestures during social interactions.

From numerous existing speech-driven face-expression and body-gesture generative models, we select representative methods based on their state-of-the-art performance metrics, such as realism, Fréchet Gesture Distance, and beat alignment. For 3D representation, we use the SMPL-X parametric model [Pavlakos et al. 2019]. The AMUSE model [Chhatre et al. 2024] emphasizes emotional 3D body gestures; we combined AMUSE with FaceFormer [Fan et al. 2021], a SMPL-X-compatible, speech-driven face animation model. Additionally, we included a real-human baseline by capturing a performer's face and body using PIXIE [Feng et al. 2021a], which incorporates DECA [Feng et al. 2021b] to predict detailed 3D facial displacements. Using AMUSE+FaceFormer (speech-driven) and PIXIE+DECA (video-driven), we evaluated perceived congruency between facial animations and body gestures.

2 Implementation Details

We represent 3D geometry using the SMPL-X model [Pavlakos et al. 2019] as a mesh $M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi})$ encoding identity shape ($\boldsymbol{\beta} \in \mathbb{R}^{300}$), pose ($\boldsymbol{\theta} \in \mathbb{R}^{J\times 3}$), and facial expressions ($\boldsymbol{\psi} \in \mathbb{R}^{100}$).

Dialogue is generated via templated responses to scenario-based questions and converted into emotional speech using PlayHT TTS[1], configured for an adult male storytelling style with neutral tempo and loudness.

FaceFormer and AMUSE differ in input processing and model architectures: FaceFormer extracts speech features using Wav2Vec and employs an autoregressive transformer [Vaswani 2017] for facial expression synthesis, while AMUSE utilizes a Vision Transformer (ViT) [Dosovitskiy 2020] to disentangle content-, emotion-, and style-related features, explicitly modeling emotional influence on gestures via conditional latent diffusion [Rombach et al. 2022]. Both generate 3D animations directly from raw audio; resulting $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ parameters are imported into Blender using the SMPL-X add-on, enabling rigged animation with blend shapes (see fig. 1, bottom left).

To evaluate synthetic animations against real human motion, we also utilize a video-based regression pipeline. PIXIE [Feng et al. 2021a] estimates pose $\boldsymbol{\theta}$, expression $\boldsymbol{\psi}$, gender-specific shape $\boldsymbol{\beta}$, and albedo $\boldsymbol{\alpha}$, while DECA [Feng et al. 2021b] extracts high-fidelity facial displacements. We record an actor responding to scenario-based queries, extracting and processing video frames with PIXIE and DECA to reconstruct geometry, albedo, and lighting parameters. Original audio synchronizes lip movements. Detailed UV-mapped shaded textures, including 3D displacements, are applied frame-wise and exported as Wavefront OBJ files. Blender's Geometry Nodes editor instantiates and animates these objects, creating mesh sequences (see fig. 1, top left). All animations share an identical outdoor environment background.

[1]https://play.ht/

3 Evaluation

RQ. "How does congruency between facial expressions and body gestures affect participants' preferences for full-body co-speech animation?" We assess if users prefer synchronized facial and body animations over those featuring a single active modality. Prior research indicates that face and body perception [Simhi and Yovel 2020] and synchronized movements improve realism and interaction quality [Fraser et al. 2022].

Study design. In a within-subjects experiment, participants interacted with three avatars in a neutral emotion scenario: body gestures only (with jaw rotations), facial expressions only (with idle body), and both modalities combined. The avatar responded to the topic *"Weekend plans"* with: *"Cooking a comforting meal becomes a therapeutic experience on lazy Sundays, as I experiment with recipes, savoring the joy of creating something delicious."* Participants selected their preferred animation mode: combined face and body, body-only, face-only, or none.

Apparatus. We used an HTC VIVE Pro 2 headset (90 Hz, 120° FOV, 2448×2448 per-eye resolution) with integrated headphones. Two SteamVR 2.0 base stations tracked participant positions. The virtual environment was built in Blender 3.4 with OpenXR-based SteamVR integration, rendering at 30 FPS on a PC (NVIDIA RTX A6000).

Procedure. Participants received an introduction and provided written consent. Wearing the HMD, they faced a virtual character positioned 1.5 m away, allowing comfortable eye contact. After experiencing the interaction scenario, participants removed the headset and completed a scenario-specific survey.

4 Results

Forty-eight participants (28 male, 20 female; ages 19–48, $\mu = 26.71$, SD=5.30) evaluated the congruency of animations with speech. Overall, 88.54% preferred combined facial and body animations. For the video-based method, 87.5% (42 participants) chose combined animation, with 6.25% (3 each) preferring body-only or face-only modes. In the audio-driven method, 89.58% (43 participants) selected combined animations, 10.41% (5) preferred body-only, and none selected face-only. Both methods exhibited high congruency between gestures and speech. The strong preference for combined modalities highlights that congruent facial and body animations enhance perceived realism. Our findings suggest incorporating VR-based subjective metrics, like perceived congruency, into model training can align outputs more closely with user perception. Iterative immersive user feedback could further refine real-time synchronization, enhancing virtual assistant responsiveness and engagement.

References

Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. 2023. Listen, Denoise, Action! Audio-Driven Motion Synthesis with Diffusion Models. *ACM Trans. Graph.* 42, 4 (2023), 1–20. doi:10.1145/3592458

Aggelina Chatziagapi, Louis-Philippe Morency, Hongyu Gong, Michael Zollhöfer, Dimitris Samaras, and Alexander Richard. 2025. AV-Flow: Transforming Text to Audio-Visual Human-like Interactions. *arXiv preprint arXiv:2502.13133* (2025).

Kiran Chhatre, Radek Daněček, Nikos Athanasiou, Giorgio Becherini, Christopher Peters, Michael J. Black, and Timo Bolkart. 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).* 1942–1953. https://amuse.is.tue.mpg.de

Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J. Black. 2019. Capture, Learning, and Synthesis of 3D Speaking Styles. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019.* Computer Vision Foundation / IEEE, 10101–10111. doi:10.1109/CVPR.2019.01034

Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. 2023. Emotional Speech-Driven Animation with Content-Emotion Disentanglement. ACM. doi:10.1145/3610548.3618183

Anna Deichler, Jonas Beskow, and Axel Wiebe Werner. 2024. Gesture Evaluation in Virtual Reality. In *GENEA: Generation and Evaluation of Non-verbal Behaviour for Embodied Agents Workshop 2024.* https://openreview.net/forum?id=1G4fIPocY2

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. 2021. FaceFormer: Speech-Driven 3D Facial Animation with Transformers. *arXiv preprint arXiv:2112.05329* (2021).

Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. 2022. 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.* IEEE, 18749–18758. doi:10.1109/CVPR52688.2022.01821

Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. 2021a. Collaborative Regression of Expressive Bodies using Moderation. In *International Conference on 3D Vision (3DV).*

Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. 2021b. Learning an Animatable Detailed 3D Face Model from In-the-Wild Images. *ACM Transactions on Graphics (ToG), Proc. SIGGRAPH* 40, 4 (Aug. 2021), 88:1–88:13.

Alan Fraser, Isabella Branson, Ross Hollett, Craig Speelman, and Shane Rogers. 2022. 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 (12 2022), 981400. doi:10.3389/frvir.2022.981400

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

Ikhsanul Habibie, Mohamed A. Elgharib, Kripasindhu Sarkar, Ahsan Abdullah, Simbarashe Linval Nyatsanga, Michael Neff, and Christian Theobalt. 2022. A Motion Matching-based Framework for Controllable Gesture Synthesis from Speech. *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)* (2022).

Taras Kucherenko, Rajmund Nagy, Youngwoo Yoon, Jieyeon Woo, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2023. The GENEA Challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In *Proceedings of the 25th International Conference on Multimodal Interaction* (Paris, France) *(ICMI '23).* Association for Computing Machinery, New York, NY, USA, 792–801. doi:10.1145/3577190.3616120

Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Linchao Bao, and Zhenyu He. 2023. Audio2Gestures: Generating Diverse Gestures from Audio. arXiv:2301.06690 [cs.CV] https://arxiv.org/abs/2301.06690

Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. 2021. AI Choreographer: Music Conditioned 3D Dance Generation with AIST++. arXiv:2101.08779 [cs.CV] https://arxiv.org/abs/2101.08779

Yuanzhi Liang, Qianyu Feng, Linchao Zhu, Li Hu, Pan Pan, and Yi Yang. 2022. SEEG: Semantic Energized Co-speech Gesture Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J. Black. 2024. EMAGE: Towards Unified Holistic Co-Speech Gesture Generation via Expressive Masked Audio Gesture Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 1144–1154.

Muhammad Hamza Mughal, Rishabh Dabral, Ikhsanul Habibie, Lucia Donatelli, Marc Habermann, and Christian Theobalt. 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).* 1388–1398.

Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. 2024. From Audio to Photoreal Embodiment: Synthesizing Humans in Conversations. In *IEEE Conference on Computer Vision and Pattern Recognition.*

Miles L. Patterson, Alan J. Fridlund, and Carlos Crivelli. 2023. Four Misconceptions About Nonverbal Communication. *Perspectives on Psychological Science* 18, 6 (2023), 1388–1411. doi:10.1177/17456916221148142 arXiv:https://doi.org/10.1177/17456916221148142 PMID: 36791676.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).* 10975–10985.

Hai Xuan Pham, Yuting Wang, and Vladimir Pavlovic. 2017. End-to-end Learning for 3D Facial Animation from Raw Waveforms of Speech. *CoRR* abs/1710.00920 (2017). arXiv:1710.00920 http://arxiv.org/abs/1710.00920

Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. 2021. MeshTalk: 3D Face Animation from Speech using Cross-Modality Disentanglement. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021.* IEEE, 1153–1162. doi:10.1109/ICCV48922.2021.00121

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 10684–10695.

Felix Sharkov, V. Silkin, and O. Kireeva. 2022. Non-verbal signs of personality: Communicative meanings of facial expressions. *RUDN Journal of Sociology* 22 (06 2022), 387–403. doi:10.22363/2313-2272-2022-22-2-387-403

Mingyi Shi, Dafei Qin, Leo Ho, Zhouyingcheng Liao, Yinghao Huang, Junichi Yamagishi, and Taku Komura. 2024. It Takes Two: Real-time Co-Speech Two-person's Interaction Generation via Reactive Auto-regressive Diffusion Model. arXiv:2412.02419 [cs.SD] https://arxiv.org/abs/2412.02419

Noa Simhi and Galit Yovel. 2020. Independent contributions of the face, body, and gait to the representation of the whole person. *Attention, Perception, & Psychophysics* 83 (10 2020), 1–16. doi:10.3758/s13414-020-02110-2

Chloe Stewart, Derek Mitchell, Stephen Pasternak, Paul Tremblay, and Elizabeth Finger. 2024. The nonverbal expression of guilt in healthy adults. *Scientific Reports* 14 (05 2024). doi:10.1038/s41598-024-60980-0

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. 2023. CodeTalker: Speech-Driven 3D Facial Animation with Discrete Motion Prior. (2023), 12780–12790. doi:10.1109/CVPR52729.2023.01229

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

Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. 2023. Generating Holistic 3D Human Motion from Speech. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 469–480.

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity. *ACM Transactions on Graphics* 39, 6 (2020).