Towards Reliable Human Evaluations in Gesture Generation: Insights from a Community-Driven State-of-the-Art Benchmark

Supplementary Material

A. Additional Results

Since this paper was submitted, we have extended our benchmark with an additional model from the authors of the Seamless Interaction dataset [2]. The Seamless model is a 250M-parameter Diffusion Transformer [62] trained with a conditional flow-matching objective on the Seamless Interaction dataset, which we evaluated in a zero-shot generalisation setting using the BEAT2 test set under the same conditions as before. As part of this extended evaluation, we also collected new votes for previous systems and system pairs, altogether extending our public database of human preference ratings from 16,000 to 20,000 votes.

We present the extended evaluation results in Fig. 6 and Fig. 7. To our surprise, the Seamless model not only matched BEAT2 motion capture in terms of motion realism, but also achieved near-indistinguishable appropriateness scores in our audio-mismatching evaluation. This is a breakthrough result that underlines the importance of evolving datasets and standardised evaluations for tracking the state-of-the-art. We emphasise, however, that these high ratings do not imply that the generated motion is perfect: the BEAT2 dataset itself contains motion-capture artifacts, and its actors exhibit varying levels of expressivity and alignment in their gestures. Similarly, our mismatching evaluation is just a first (albeit important) step towards measuring multimodal alignment: there is a strong need for reliable evaluation methods for measuring, e.g., semantic or emotional alignment of gestures.

B. Surveyed Publications

In Tab. 1 we present the list of 26 publications surveyed in Sec. 3. This selection is a result of searching for the terms "gesture", "co-speech", "speech", and "motion" in publication titles at CVPR, ICCV, ECCV, SIGGRAPH, and SIGGRAPH Asia, then filtering down the results to models whose outputs include 3D body gesture. While the selected conferences do not cover all of gesture generation research, we believe our survey gives a representative picture of the evaluation practices for state-of-the-art models.

Looking at the second-to-last column of Tab. 1, we note the worrying trend of recent publications not comparing against the reference human gestures (from motion capture). In the last column, we show all direct comparisons amongst the surveyed works driven by human evaluations (cf. Sec. 3.2): each **M** entry is indicates that the model in that row was compared to the baseline model **B** in that column. Overall, the lack of direct comparisons against human motion and strong baseline models makes it challenging to assess the progress made by most publications.

C. Additional Details on our Evaluation Protocol

C.1. Test Segment Selection

Following prior large-scale evaluations in speech [63] and gestures [42–44], we use short speech-gesture segments as the stimuli for our user studies. Using this one set of speech segments as the basis for all user studies carried out on the BEAT2 dataset, we control for the effect of the speech on system behaviour (e.g., the same speakers are always represented in the same way in every evaluation).

In particular, we curate 108 evaluation segments from the BEAT2 English test set, covering all speakers. The segments are randomly sampled complete sentences, manually filtered for artifacts like flickering and self-intersection. We selected four segments for most speakers, and eight for Scott and Wayne due to their higher mocap quality. This larger segment count, compared to typical evaluations (e.g., GENEA Challenges) allows for more reliable user studies whilst leaving room for analysis on the stimulus level. The criteria for selecting speech segments for the user studies were as follows:

- Each segment should correspond to one or more complete sentences.
- Segment duration should be within the range of 7.0 to 12.0 seconds.
- Segments should be disjunct (no overlap).
- Finally, the BEAT2 SMPL-X motion capture for the segments should not contain any major artefacts.

The use of complete sentences is more pleasing to test-takers and means that every segment starts and ends at a sentence boundary. Sentences were identified automatically using the text transcription provided by the dataset. The specific duration range was chosen based on an informal evaluation by paper authors on all test-set speakers, which indicated that segments shorter than seven seconds too often contained no gesturing at all, whereas segments longer than twelve seconds were difficult to pay sufficiently close attention to throughout, or varied more in quality to the extent that they were more difficult to assign a rating to.

Although a lot of the test dataset contains somewhat awkward finger poses due to the difficulty of tracking fingers, this was deemed less visually distracting and would

Table 1. Overview of human evaluation practices in 3D gesture-generation research published at SIGGRAPH, SIGGRAPH Asia, and leading computer-vision venues between 2023–2025, as described in Sec. 3. The table uncovers the fragmented state of human evaluation, with inconsistent study designs for related tasks (*Tasks* column), and a critically low degree of direct comparisons between top models (last column). Abbreviations: SG=SIGGRAPH; **Na**=Naturalness; **Re**=Realism, Plausability or Believability; **Hu**=Human-likeness; **Sm**=Smoothness; **Pref**=Preference; **Rh**=Rhythmic; **Sem**=Semantic;**Gen**=General; **Em**=Emotion; **St**=Style; **B**=Present in direct comparison as baseline; and **M**=Present in direct comparison as main model.

Year	Venue	Model	Training dataset	Modelling Goal		Directly compared to...	
				Quality	Alignment	Mocap	A model in the survey
2023	CVPR	DiffGesture [90]	TED [83], TED-Expr. [90]	Na, Sm	Rh	✔	B
		QPGesture [79]	BEAT [47]	Hu	Gen	✔	B
		RACER [69]	Trinity [29], own	Re	Gen, Sem	✗	
		TalkSHOW [82]	SHOW [82]	✗	Gen	✔	B
	SG.	Bodyformer [60]	Trinity [29], TWH [45]	Hu	Gen	✔	
		GestureDiffuCLIP [7]	BEAT [47], ZEGGS [30]	Hu	Sem, St	✔	B
		LDA [6]	Trinity [29], ZEGGS [30]	Pref	St	✔	B
	ICCV	C-DiffGAN [4]	PATS [3]	Na	Sem, Rh, St	✔	
		LivelySpeaker [89]	BEAT [47], TED [83]	Na, Sm	Sem	✗	
2024	CVPR	AMUSE [19]	BEAT2 [48]	✗	Rh, Em	✔	M
		Audio2Photoreal [58]	own	Re	✗	✔	M
		ConvoFusion [56]	DnD [56]	Na	Gen, Sem	✔	
		DiffSHEG [15]	BEAT [47], SHOW [82]	Re	Rh	✗	M M B
		EMAGE [48]	BEAT2 [48]	Re	✗	✗	M B
		EmoTransition [65]	own	Na, Sm	✗	✗	M
		ProbTalk [53]	SHOW [82]	✗	✗	✗	B
	SG.	Sem. Gest. [87]	BEAT [47], ZEGGS [30], own	Hu	Sem, Rh	✔	M
	SG. Asia	SIGGesture [17]	BEAT [47]	Na	Sem, Rh	✗	M M M
2025	CVPR	HOP [16]	TED [83], TED-Expr. [90]	Na, Sm	Sem, Rh	✔	
		LOM [14]	BEAT2 [48]	✗	✗	✗	
		RAG-Gesture [57]	BEAT2 [48]	Na	Sem	✔	M
	SG.	MeCo [13]	BEAT2 [48], ZEGGS [30]	Hu	Gen	✗	M
	ICCV	GestureHydra [77]	SHOW [82], own	Na	Sem	✗	M M
		GestureLSM [50]	BEAT2 [48]	Sm, Re	Rh	✗	M M M
		SemGes [49]	BEAT [47], TED-Expr. [90]	Na	Sem, Rh	✔	M
		SemTalk [86]	BEAT2 [48], SHOW [82]	Re	Sem, Rh	✗	M

be more drastic to exclude, so it was not considered grounds for exclusion. The only cases where mesh penetration were permitted were (2a) when the penetration visually resembled clothing or tissue giving way to light pressure, or (2b) where the penetration occurred due to the aforementioned poor finger tracking, as long as the fingers at worst were merely seen clipping into each other, and not passing through each other to the other side at any point. We decided to retain segments satisfying (2b) for the evaluations because these instances of mesh penetration are associated with poses having gestural importance, such as interwoven hands or finger against palm, which are important not to exclude.

After the segment selection was performed, we measured the potential rhythm bias of our alignment evaluation by comparing our evaluation segments to the semantic labels of BEAT2; we found that $\approx 59\%$ of segments fully contain at least one semantic gesture, even though we did not intentionally select segments with semantic annotations. In other words, our benchmark heavily oversamples semantic gestures compared to their natural frequency, and is not unnecessarily unfair towards models aiming at semantic generation.

Overall, gesture generation aims for realistic and expressive animation rather than replication of the dataset. This distinction is important due to pose estimation artifacts and the natural variation of human expression, and necessitates careful selection of evaluation segments. Overlooking this step, as most evaluations in our survey do, can lead to lower scores for the reference human motion, and ultimately, im-

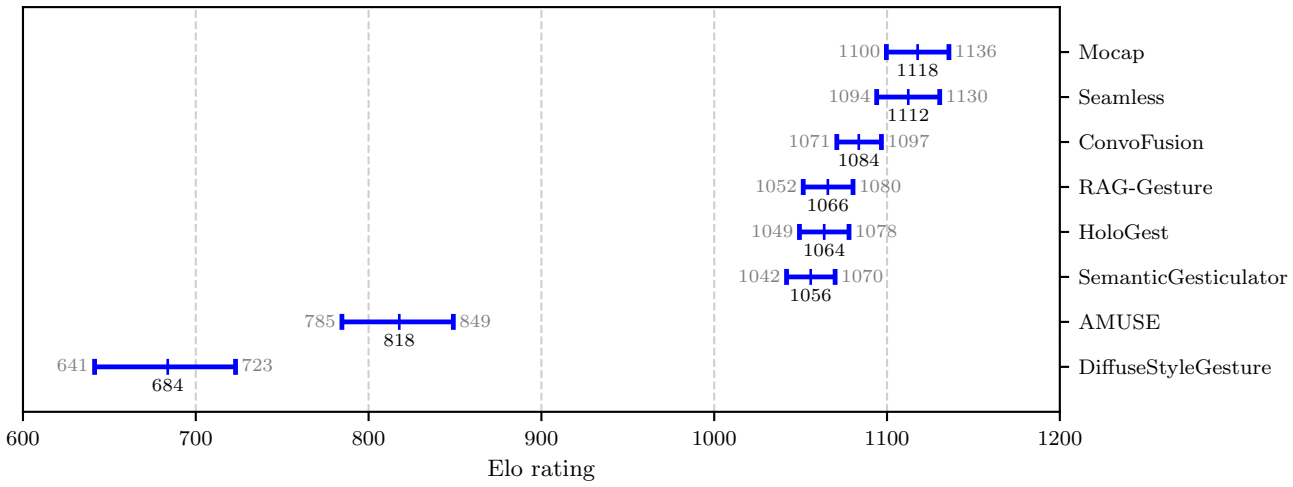

Figure 6. Updated Elo ratings after adding the Seamless model to the benchmark.

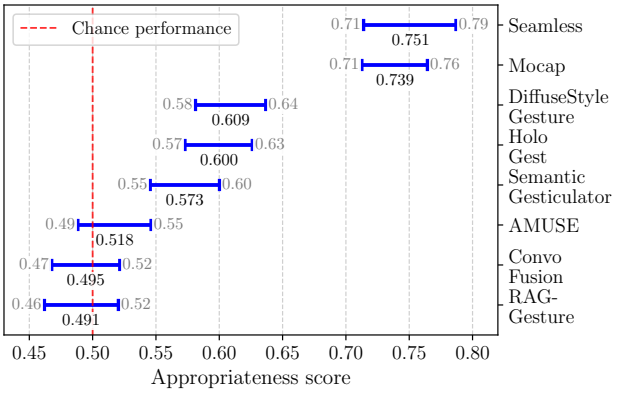

Figure 7. Updated appropriateness scores after adding the Seamless model to the benchmark.

Figure 8. A video frame showing a gesturing SMPL-X avatar (male variant) rendered using the Blender visualiser.

precise results.

C.2. 3D Visualisation

Visualisation plays a critical role in the evaluation of gestural motion. Prior work has demonstrated that the quality and type of visualisation can significantly impact the outcome of evaluations [58]. Therefore, it is essential to standardise the visualisation pipeline to ensure that comparisons across systems are made on equal terms. In line with the direction of the field towards increasingly photorealistic avatars [12, 58], we aim to provide high-quality, realistic visualisation. High-quality renderings have been found to aid human raters in distinguishing between better and worse motion, providing clearer, more consistent evaluation results [58].

Following the approach of the GENEA Challenges [44, 85], our visualisation (Fig. 8) includes full-body motion with root-node translation and rotation. This captures the character's positioning and stance with respect to the camera, leading to a more lifelike and expressive visualisation than methods that restrict animation to only the upper body [19, 42, 60, 87].

Since BEAT2 is in the SMPL-X format, we use SMPL-X meshes for our visualisation, which offer an anatomically accurate body shape and proportion that matches each speaker [61]. Although SMPL-X includes gendered meshes, we opted to use the gender-neutral mesh for all characters, which is further modified by the per-speaker body shape parameters. This decision was made because the gendered meshes often resulted in increased self-intersection when visualising motion from the BEAT2 test set. However, the gender-neutral mesh lacks features such as hair and clothing and does not replicate realis-

tic skin tones or facial characteristics. The male- and female-presenting speakers are distinguished through different SMPL-X textures: male characters use the default white shirt texture, and female characters use the pink one, matched to the perceived gender of the voice. These textures also introduce variation in skin tone. To enhance realism, we added a hair prop and applied a displacement map to the clothing, improving its appearance and reducing the flatness seen in the SMPL-X mesh. Due to the absence of gaze information and inaccurate lip sync in BEAT2, we covered the face with a white mask. This avoids distracting visual artefacts that could negatively influence test-taker perception.

Camera positioning was carefully determined to match the approximate viewpoint of a listener being addressed by the speaker. To achieve this, we calculated the mean root-node (hip) translation over the animation frames and used it to normalise the speaker's position in the scene. The camera was then statically placed based on this average position to ensure consistent framing across clips. This setup allows for full visibility of the speaker's hand and arm gestures, though depending on the magnitude of motion, speakers may appear slightly closer to or further from the camera. In rare cases, if the root translation is particularly large, the character may briefly move out of frame.

Furthermore, we excluded the feet from view, cutting the frame at approximately knee level. This was a deliberate choice motivated by known issues with foot-ground interactions, such as sliding and ground penetration, which are both common in synthetic motion and easily detectable by human observers. If shown, these artefacts would likely overshadow the gestural qualities under evaluation, as crowdsourced participants tend to focus on the most visually salient errors (see, for instance, the results in Appendix D.3). By omitting the feet, we help raters concentrate on the gestures themselves, aligning with the primary evaluation objective.

It is also worth noting that while foot-ground contact issues are primarily governed by straightforward laws of physics rules and can potentially be resolved through post-processing, upper-body gesturing is much less dictated by physical laws and is instead rooted in communicative and cultural conventions. This arguably makes gesturing a deeper and more challenging problem to solve in the long term, compared to issues of character interaction with the ground plane. For the record, while the human evaluations are conducted based on cropped visualisations, the automatic metrics described in Appendix F operate on the full-body pose and motion data.

The rendering environment was kept neutral, using only ambient lighting without shadows. This setup speeds up rendering while preserving sufficient visual fidelity. A simple indoor background was chosen to minimize distractions and keep the viewer's attention on the animated character.

C.3. User-Study Setup

For the user-study screens, we propose the layout and phrasing (for instructions and response options) shown in Fig. 9. Our implementation of this interface, used for the experiments in Sec. 5, is illustrated in Fig. 10. We suggest presenting each crowdsourced participant with 25 screens, leading to 25 pairwise votes collected in total (barring technical issues).

C.3.1. Details on Participant Recruitment

As participant recruitment may highly depend between evaluation setups, we do not aim to standardise it in our protocol. Regardless, we share important details from our evaluations in Sec. 5 below.

Test-takers are recruited through the Prolific crowdsourcing platform. To be eligible to participate, they are required to reside in any of six English-speaking countries (Australia, Canada, Ireland, New Zealand, the United Kingdom, and the USA) and to have English as their first language. The number of participants and their demographics can be found in Tab. 2. No Prolifc user is allowed to participate in the same user study more than once, although this constraint is not enforced between different user studies. Remuneration is set at with 5.25 GBP for a successfully completed test, corresponding to a median of 12.6 GBP hourly rate quoted by the Living Wage Foundation in the UK, computed from the approximate study duration of 25 minutes.

C.3.2. Attention Checks

Consistent with best practices in crowdsourced evaluations, our protocol includes attention checks to ensure that test-takers are paying attention to the task. These take the form of a message "[Attention check] Please choose '*R*'.", with *R* being one of the five response options underneath the videos, chosen at random. Four attention checks are inserted into each user study, evenly spaced from the 20% until the 80% progress mark. Test-takers that fail any attention check are removed from the statistical analyses; those that fail more than one are rejected without pay. (Prolific's policies do not permit rejecting test takers due to a single failed attention check.)

For the realism evaluation, the attention-check message is presented as high-contrast, easy-to-read text superimposed on one of two otherwise normal video stimuli in a pair. For the speech appropriateness evaluation, each test taker is subjected to two visual (text-based) attention-checks as above, along with two audio attention checks, in which the video is unaffected but the speaker audio in one of the videos is partly replaced by a synthetic voice speaking the same message. In all cases, attention-check messages do not appear until a few seconds into each attention-check

Table 2. Demographic statistics of crowdsourced test takers that participated in our motion realism evaluation and our mismatching study for speech-gesture appropriation. The age is given as an average and a standard deviation.

Study	Test-takers	Country of residence						Sex			Age (years)
		US	UK	CA	AUS	IE	NZ	M	F	N/A	
Realism	336	202	72	48	10	2	2	198	137	1	39 ± 12
Approp.	311	154	123	19	8	3	2	175	132	3	39 ± 13

video, so that test-takers who only pay attention the first seconds are likely to fail the checks. All test-taker responses given in response to attention checks is excluded from the statistical analyses. Finally, in the case of technical errors such as videos not loading, participants may skip up to three study screens; when a fourth skip occurs, the study is terminated, and a manual review is triggered to establish whether the participant should be paid. This means that each test taker who successfully completes a user study contributes between 23–25 total responses (comprising a one of five possible preference indications and the associated responses to the JUICE questions).

D. Additional Details on Our Benchmarking

D.1. Statistical Analysis for Motion Realism

Our statistical analysis is based on the pairwise preference data acquired from the evaluations. We standardise the condition labels to canonical forms and convert the raw choices into triplets of the form $(\text{model}_A, \text{model}_B, \text{winner})$, as required by our scoring algorithm. A "clear" preference counts as two wins for the winner, whereas a "slight" preference only counts as one; ties count as half a win and half a loss for both models in the presentation.

To transform the pairwise preferences into a continuous ranking, we use the Bradley-Terry Elo-style model advocated by the Chatbot Arena team [20]. This approach preserves the interpretability of classical Elo ratings while avoiding the dependence on update order, which can distort results in online systems with large K values. Specifically, we consider the latent skill of each system as a real-valued parameter e and postulate that for any pair (A, B) the log-odds of A beating B are equal to $e_A - e_B$ divided by a scale constant. Under a logistic link, this assumption yields the Bradley-Terry probability [10], which is maximized in a single-batch optimization rather than incrementally. Following generally accepted standards of Elo calculation we set the scale to 400, so that a 200-point difference corresponds to 76% probability of winning. This reflects practices in the game of chess, for example. The model also assumes that the maximum likelihood estimates of the ratings are approximately Gaussian when the number of pairwise comparisons is large, allowing for easy calculation of standard errors. We exploit this asymptotic normality to derive Wald confidence intervals for each rating and to propagate uncertainty when computing derived quantities such as predicted win rates. Because the pair frequencies are unbalanced, we additionally perform non-parametric bootstrapping over the original trials to guard against violations of the Gaussian approximation. In each bootstrap replicate, we sample battles with replacement, fit the Bradley-Terry model, and record the resulting set of Elo ratings. All optimization is done using `scikit-learn`'s unconstrained logistic regression solver, which reliably converges to our dataset within seconds.

D.2. Statistical Analysis of Speech Appropriateness

For the statistical analysis, we use the basically same setup as described in Appendix D.1 for motion realism: clear preference responses count double compared to slight preferences, and ties ("They are equal") count as half a win and half a loss. The only difference is that wins for the matched stimulus are assigned the value 1 and wins for the mismatched stimulus are assigned a 0. The resulting average *appropriateness score* (essentially a modified win rate) is then a number between zero and one.

The rest of the analysis is the same as for motion realism. We use the exact same test-taker-level bootstrap methodology to obtain confidence intervals, based on quantiles of the bootstrap distribution.

D.3. JUICE Scores for Motion Realism

Although we collect JUICE responses for all presentations where there was not a tie, we here focus on analysing the JUICE responses when comparing each of the six initial synthetic systems to the mocap topline. The distribution of these is graphed as a bar chart in Fig. 11. (All JUICE responses, including free text for the "Other" option, are featured in the data we release.)

Fig. 11 was created by, for each system, first counting how often each of the five JUICE options (tickboxes) were ticked when that system was pitted against the BEAT2 mocap. (Ties, i.e., "They are equal" are ignored since they did not generate JUICE responses.) After that, we normalised these values into percentages, where 100% would mean that the specific JUICE option in question was ticked every single one of these presentations. Finally, we split the percentages by whether or not they were associated with a win (the bar pointing up from zero) or a loss (the same bar extending down below zero instead) for the system in question. Our normalisation brings forth the qualitative profile of a model by compensating for imbalances in the total number of responses, which vary because comparisons with large perceptual differences pause early. Furthermore, strong wins and losses were counted as one win or loss instead of two

Below are two videos without audio of a character speaking and gesturing.

Left video | Right video

In which video does the character gesture more like a real person?

Left clearly better | Left slightly better | They are equal | Right slightly better | Right clearly better

Which factors contributed most to your response? Please tick one or more options:

☐ Unrealistic motion (glitches/artefacts, limbs/body penetrating each other, physically impossible motion)
☐ The smoothness of the motion
☐ The amount and intensity of motion
☐ Recognisable gestures
☐ Other (Please specify factors not listed above): _____

Below are two videos of a character speaking and gesturing. Both videos have the same motion, but different speech.

Left video | Right video

In which video do the character's movements fit the speech better?

Left clearly better | Left slightly better | They are equal | Right slightly better | Right clearly better

Which factors contributed most to your response? Please tick one or more options:

☐ Fit the rhythm and timing of the speech better
☐ Emphasised the correct part (or parts) of the speech
☐ Better matched the content and meaning of the speech
☐ Better fit for the emotion of the speech
☐ Other (Please specify factors not listed above): _____

Figure 9. Questions and response options in the two types of user studies, also showing their schematic layout in the user-study GUI. For a screenshot of the GUI see Fig. 10.

in our analyses of JUICE responses in this paper. Together, this setup means that the percentages in the plot are biassed towards factors that correspond to subtle differences.

Across the top four systems, the *Smoothness of the motion*, the *Amount and intensity of the motion*, and *Recognisable gestures* were each ticked at comparable rates. This broad similarity mirrors the tight Elo clustering observed earlier and suggests that evaluators focus on nuanced aspects of kinematics when the realism gap to the motion capture is small. The catch-all category *Other reason* was used relatively sparingly for every system, suggesting that the four predefined options captured most salient perceptual differences. Analysis of the free-text responses is left as future work.

The most notable deviation from the general uniformity of response rates to pre-defined JUICE options is the frequency with which the *Unrealistic motion* option was chosen. Although selected less often, it is disproportionately associated with AMUSE and especially DiffuseStyleGesture, the two systems that occupy the lower end of the Elo ratings in this study. This clear pattern supports an interpretation that visible artefacts, such as jerks, implausible limb trajectories, and temporal discontinuities, are a primary cause of dispreference when present and must be not be generated for competitive performance.

Past gesture-synthesis systems have been criticised for producing "marginally natural gestures that appear more like well-timed hand waving, are not communicative and have little meaning" [59]. As such, it might a-priori be expected that synthetic systems may struggle to produce distinctive and recognisable communicative gestures, e.g., iconic and metaphoric gestures. We therefore find it surprising to see that "recognisable gestures" did not show any apparent advantage for mocap over synthetic gestures. Although it is possible that strong contemporary systems have improved on the issues pointed out by Nyatsanga et al. [59], e.g., with the RAG-Gesture system [57] employing retrieval-augmented generation, it is also possible that this option might need to be replaced by another formulation and/or be complemented by additional instructions in the

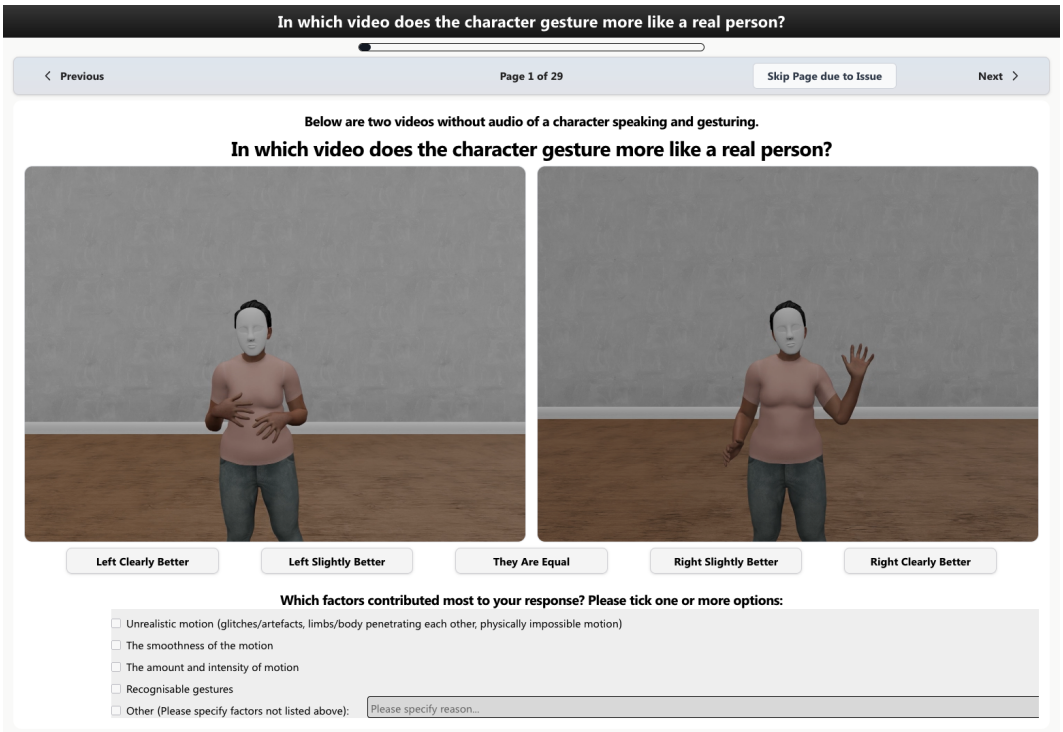

Figure 10. A screenshot of the GUI for the user studies, specifically from a motion-realism test with the current screen containing stimulus videos of the female avatar. The JUICE options are disabled with a grey background since no response to the main question has been selected yet.

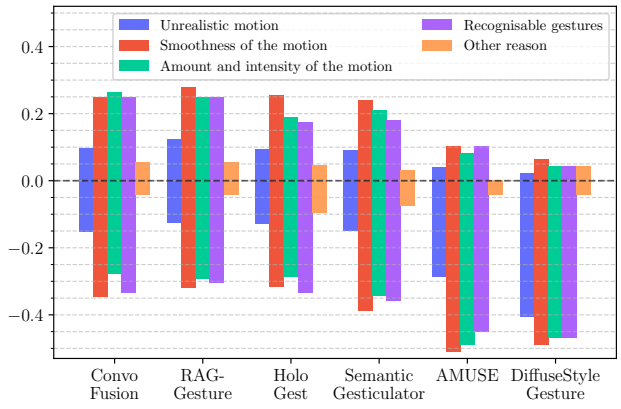

Figure 11. Frequency of JUICE options chosen for each model during the motion-realism evaluation in comparisons to the motion-capture condition, ignoring ties. Bar plots above zero show frequency among winning outcomes; bar plots below zero correspond to frequencies among losing outcomes, both relative to the total number of non-tie comparisons for the given model.

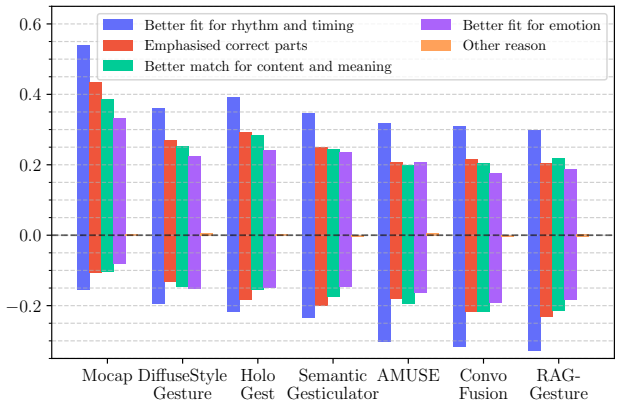

Figure 12. Frequency of JUICE responses chosen in the speech-gesture appropriateness study, for each model, when compared against its mismatched counterpart. Positive values are the frequency among winning outcomes; negative values correspond to the frequency among losing outcomes, both relative to the total number of non-tie comparisons.

future that more clearly communicate its intention to crowd-sourced test takers.

D.4. JUICE Scores for Speech-Gesture Appropriateness

The distribution of normalised JUICE responses for gesture–speech appropriateness is shown in Fig. 12. Like for

the realism analysis, the number of responses for each option were first accumulated for every condition, regardless of the outcome of the trial. Then, the numbers were converted to proportions that sum to one within each condition. However, whereas the earlier Fig. 11 specifically graphed the data from cases where test takers compared clips from each artificial system to BEAT2, the appropriateness study never asks test takers to compare conditions directly, but only to assess matched and mismatched video clips within each condition. For this reason, the Mocap condition is included in Fig. 12 but not in the earlier Fig. 11.

Across all conditions, the reason *Fit the rhythm and timing of the speech* was selected disproportionally often, both when a model was preferred and when it was dispreferred, indicating that temporal alignment is most salient feature for test-taker decisions. This makes sense, given that the speech segments evaluated were not selected to contain rich semantic grounding or strong emotional colouring, making it so that rhythm naturally becomes the primary distinguishing factor. (See also Saund and Marsella [68].)

The other JUICE options *Emphasised the correct part of the speech*, *Better matched the content and meaning of the speech*, and *Better fit for the emotion of the speech*, were each chosen at similar rates and significantly less often than rhythm, but were still used an appreciable faction of the time. This implies that participants considered these aspects overall less consistently important for their choice. The catch-all category *Other reason* was used even more rarely than in the motion-realism JUICE response data, indicating that the predefined options well capture the most important sources of preference in the appropriateness domain.

Unlike the *Unrealistic motion* option in Fig. 11, there are no strong indications that certain JUICE options are selected disproportionally often for certain systems, including, say, for RAG-Gesture and Semantic Gesticulator, both of which specifically target improved semantic consistency in their work. Appropriateness evaluations with segments selected to contain semantic gesturing, or to mismatch between emotions, might alter this balance and could be interesting future work.

E. Details on Systems Evaluated

In this section, we describe the notable features of each of the six gesture-generation systems evaluated in the main paper, as well as the adaptation steps performed by each model's original authors when preparing their submissions.

E.1. DiffuseStyleGesture [78]

DiffuseStyleGesture aims to generates high-quality, speech-synchronized 3D co-speech gestures through a diffusion model architecture. The generation process ensures outputs exhibit robust temporal audio-gesture synchronization and stable kinematics. A notable feature is the incorporation of seed gestures for initialisation, providing control over the generation process, leading to varied and contextually relevant motion.

DiffuseStyleGesture was originally trained on the ZEGGS dataset [30]. To prepare the submission, the authors of the model processed input motion into a comprehensive per-joint feature set [80]. Additionally, a key modification to the published model is the use of a data filtering strategy, exclusively utilising data from a select cohort of professional actors (Actors 2, 3, 4, 7, 10, 15, 16, 17, 18, 21, and 27) to learn from high-fidelity motion exemplars. Generated joint-based positional output is converted to the SMPL-X format by mapping joint rotations (from Euler angles) to SMPL-X axis-angle pose parameters via a predefined joint map, alongside extracting and transforming the root joint's translation and coordinate system for SMPL-X alignment.

This manual conversion from positional data to SMPL-X, adopted to maintain input feature parity with [78, 80], may compromise visual fidelity compared to direct SMPL-X feature utilisation in training and generation [48]. Therefore, additional post-processing was employed in the form of a minor scaling applied to the root joint's motion to enhance visual stability and mitigate potential drift during front-facing camera evaluation, and a subtle inverse kinematics (IK) adjustment from the feet to the root, which serves as a minor refinement with negligible visual impact on foot placement.

E.2. Semantic Gesticulator [87]

Semantic Gesticulator aims to generate high-quality, semantically meaningful gesture animations from speech by combining rhythmic precision with contextual understanding. Unlike prior models that rely solely on direct audio-to-motion mappings, this model introduces a discrete latent motion space via a residual VQ-VAE, enabling compact and diverse motion representations. It uniquely integrates a GPT-based gesture generator with a large language model (LLM)-driven semantic retrieval system, which selects appropriate gestures based on transcript context. A semantics-aware alignment module then fuses rhythmic and semantic information, resulting in gestures that are both expressive and contextually appropriate.

The model's authors adapted the original system by removing the semantic gesture retrieval component and relying solely on the base RVQ+GPT pipeline for audio-to-gesture generation. This simplification allows evaluating the core generative capacity of the model. Additionally, the data preprocessing module was modified to support the SMPL-X representation used in the BEAT2 dataset, ensuring compatibility with our motion format. During training, the RVQ module was configured with a codebook size of

1024 and 4 quantization layers to accommodate the longer and more complex motion sequences present in the full BEAT2-English dataset. The GPT-based gesture generator was trained using the same architecture and settings as described in the original system. No additional postprocessing was applied to the motion output, enabling an unbiased assessment of the model's raw generation quality.

E.3. ConvoFusion [56]

ConvoFusion is a diffusion-based framework for speech- and text-driven gesture synthesis. It features a latent diffusion architecture with two components: 1) a scale-aware temporal VAE that models different body parts separately and represents sequential motion frames using temporally ordered latents, and 2) a transformer decoder for diffusion that contains separate cross-attention heads for conditioning on different modalities i.e. speech, gesture and speaker identity.

Originally, this framework was not trained on BEAT2 and is therefore modified to accommodate the SMPL-X input representation. The scale-aware VAEs are trained independently for four body parts: upper body, hands, face, and lower body. Following this, the base latent diffusion framework is adapted to the updated VAEs. Additionally, the speech representation is upgraded from mel-spectrograms to wav2Vec embeddings [9]. Leveraging the temporal structure of VAE latents, the framework is capable of auto-regressively generating long-form motion in time-windowed chunks. To perform auto-regressive rollout, it first generates the initial 10 seconds of motion, then uses the last 1 second of that output as seed motion to generate the next 9 seconds. This seed motion maintains continuity across steps through diffusion-based outpainting. At each step, overlapping motion segments are linearly blended with the previous ones, resulting in a single coherent motion sequence.

E.4. RAG-Gesture [57]

RAG-Gesture aims to generate not only natural looking but also semantically meaningful gestures. It achieves this by first training a base latent-diffusion framework for co-speech gesture generation (similar to ConvoFusion [56]), and then leveraging retrieval augmented generation during inference to inject semantically meaningful exemplars. The generated gestures are therefore sampled from the base distribution of a diffusion model, while also being semantically grounded in explicit domain knowledge, like gesture types or discourse relations. The method is agnostic to the choice of retrieval algorithm; in the original paper, two approaches were presented: one based on an LLM's understanding of gesture type, and the other grounded in discourse-based linguistic analysis of the speech.

The system is inherently trained on BEAT2 dataset and follows its input representation, therefore no adaptation to the trained model is made. Specifically it generates hand, body, face motion along with the translation of the character from a single model. As the framework follows the temporal VAE structure, it also performs long-form motion generation in chunks of 10-second time windows through autoregressive rollout (Appendix E.3). Consequently, retrieval algorithm is not used for the overlapping motion frames and RAG is performed for the newly generated motion. For evaluation, LLM-driven Gesture Type algorithm is used for RAG.

E.5. AMUSE [19]

AMUSE is an emotional, speech-driven model for 3D body animation. It converts audio filter-bank features into three disentangled latent vectors that separately encode (1) linguistic content, (2) emotional state, and (3) speaker style. The speech encoder is a Vision Transformer (ViT) [71] adapted to operate on filter-bank images. These vectors condition a latent-diffusion model [67] that generates gesture motion sequences. After training, AMUSE can synthesize 3D human gestures directly from speech while allowing users to combine content, emotion, and style, for example, pairing the content vector of a source speech with the emotion and style vectors from a different one. Stochastic sampling of the diffusion noise term yields diverse gesture variants that preserve the chosen emotional expressivity.

AMUSE was developed on the BEAT2 SMPL-X data, with the same dataset splits that we use. However, there are two differences between the submission format and the data processing of the original model that require adaptation. First, AMUSE puts emphasis on upper-body gesticulation rather than locomotion, therefore it discards the eight lower-body joints of the SMPL-X body. This was resolved by augmenting the model outputs with static lower-body joints. Second, AMUSE can only generate 10-second motion sequences, while the submission system normally expects a single, coherent motion sequence for each test-set file. As a workaround, the AMUSE submission contains 7–12 second motion clips, corresponding to the full set of speech segments described in Appendix C.1. Clips shorter than 10 seconds were generated by padding the audio input with silence, and discarding surplus motion frames from the output. Clips longer than 10 seconds were artificially created by blending two clips, c_1 and c_2, using spherical linear interpolation (SLERP), where c_1 is generated on the first 10 seconds of the segment, and c_2 is generated from the last two seconds of c_1 and the remaining portion of the segment.

E.6. HoloGest [18]

HoloGest aims to generate physically plausible and vivid co-speech gestures by addressing limitations in current diffusion-based methods, which often use a single noise dis-

tribution for full-body gestures despite their differing characteristics. It tackles this issue by decoupling body parts to learn separate noise distributions and introduces motion priors to enhance physical plausibility, effectively reducing unnatural phenomena like jitter and sliding. Additionally, HoloGest employs an adversarial generation approach to accelerate the denoising process, requiring only 50 steps (0.7 seconds) to produce 2 seconds of gestures, making it suitable for real-time performance. These strategies enable HoloGest to deliver highly realistic and dynamic gestures while maintaining computational efficiency.

The published version of HoloGest features two motion priors trained on external datasets: one for the finger motion, and another for the root trajectory. In contrast, the HoloGest submission ensures fairness towards other participating systems by removing the finger prior, and retraining the trajectory prior on the BEAT2 training set, without relying on external datasets. Furthermore, the independent diffusion generation channel for facial expressions was removed, therefore the submission only contains body- and hand motion.

F. Experiments on Automatic Metrics

We provide evaluation results using a curated set of automatic metrics, often called objective metrics, primarily selected based on their frequent use in recent gesture-generation research. While human evaluation ultimately determines overall performance, automatic metrics may serve as a complementary tool to benchmark and analyse system behaviour efficiently and at scale.

We report results of seven metrics. **Fréchet Gesture Distance (FGD)** measures the Fréchet Distance between human motion and generated motion distributions on a learnt feature space [48, 84]. **Fréchet Distance on Geometric and Kinetic Features (FD$_g$ and FD$_k$)** [6, 58]; FD$_g$ measures the Fréchet Distance between the distributions of static pose data from human and generated motion. FD$_k$, on the other hand, compares the distributions of inter-frame pose differences (i.e., motion velocity). **Beat Alignment (BA)** evaluates the alignment between the beats in the input speech and those in the generated motion [46, 51]. **Semantic Relevance Gesture Recall (SRGR)** compares human motion and generated motion by evaluating the proportion of correctly recalled joints only over segments containing semantic gestures [47]. **Pose Diversity (DIV$_{pose}$)** evaluates how diverse the generated poses are within each motion sequence by computing the average deviation of individual poses from the mean pose. **Sample Diversity (DIV$_{sample}$)** measures the diversity across multiple generated motion samples for the same input, indicating the stochastic variability of the model's outputs.

Table 3 presents the results on the test set of the BEAT2 dataset. RAG-Gesture shows strong performance for distribution-based motion-quality metrics (FGD, FD$_g$, FD$_k$), as well as BA. HoloGest shows the best SRGR and DIV$_{\text{pose}}$. Semantic Gesticulator yields the best FGD and richest run-to-run variability (DIV$_{\text{sample}}$). Note that the FGD values are not directly comparable to those in Liu et al. [48] due to differing data sizes (see Chong and Forsyth [21]): we used all audio in the test set and all five random samples submitted for each system to obtain as many data points as possible for better distribution fitting.

We examined the correlation between the results of the automatic metrics and the subjective human ratings. First, we looked at the motion-realism-related metrics, FGD, FD$_g$, and FD$_k$. According to the user study, the ConvoFusion, RAG-Gesture, HoloGest, and Semantic Gesticulator systems achieved relatively high Elo ratings, while AMUSE and DiffuseStyleGesture had lower ratings compared to the other systems. When roughly dividing the systems into these two groups, we observed that the high Elo-rating group consistently outperformed the low Elo-rating group on the FGD and FD$_g$ metrics, which aligns with the user ratings.

Next, regarding speech-gesture appropriateness, we considered the BA and SRGR metrics. Here, we found a substantial discrepancy between the automatic metrics and human ratings. For example, RAG-Gesture achieved the best BA score, but had the lowest user rating in the user study. Similarly, although HoloGest, RAG-Gesture, and ConvoFusion achieved high SRGR scores, they did not demonstrate clear appropriateness in the human evaluations.

Inspired by the GENEA Challenges [44], we also conducted a quantitative analysis of the correlation between automatic metric scores and subjective human ratings. Given the limited number of systems, the correlation analysis serves as a reference and should not be interpreted as providing strong evidence or conclusive findings. Specifically, we computed the correlation between Elo ratings (representing human preferences) and each automatic metric using Kendall's τ rank correlation [38]. For human ratings on motion realism, all motion quality metrics exhibited moderate negative correlations (between -0.4 and -0.6, consistent with the findings for FGD in Kucherenko et al. [44]), while SRGR – despite being more closely related to speech-gesture alignment – showed the highest positive correlation (0.73). For speech-motion alignment, BA demonstrated a moderate correlation (0.5), whereas SRGR showed virtually no correlation (-0.14). However, none of these correlations were statistically significant ($p < 0.05$).

Overall, our findings highlight that whilst automatic metrics can provide useful insights and facilitate early evaluation, they remain insufficient to replace human evaluation in gesture generation. The discrepancies – especially in speech–gesture alignment – underscore the limitations of current objective measures and the continued necessity of

Table 3. Automated evaluation of gesture generation models using a set of objective metrics. Human motion capture data is included for reference. The best value for each metric among systems is **boldfaced**. For BA and DIV_{pose} metrics, values closer to the human reference motion are considered better.

	Condition	FGD↓	FD_g↓	FD_k↓	BA→	SRGR↑	DIV_{pose}→	DIV_{sample}↑
	Motion capture	0.000	0.000	0.000	0.645	1.000	8.302	–
Systems	HoloGest	0.625	0.972	0.059	0.539	**0.469**	**7.733**	0.011
	RAG-Gesture	0.515	**0.660**	**0.035**	**0.648**	0.427	10.092	0.013
	DiffuseStyleGesture	7.110	10.128	0.099	0.608	0.312	9.598	0.001
	Semantic Gesticulator	**0.473**	0.749	0.043	0.681	0.398	10.993	**0.020**
	ConvoFusion	0.600	0.817	0.040	0.611	0.448	8.911	0.013
	AMUSE*	0.785	0.997	0.041	0.757	0.394	9.552	0.018

*AMUSE results are affected by motion discontinuities stemming from its lack of long sequence support.

human evaluation.