Gesture Generation (Still) Needs Improved Human Evaluation Practices: Insights from a Community-Driven State-of-the-Art Benchmark

Rajmund Nagy[1], Hendric Voss[2], Thanh Hoang-Minh[3], Mihail Tsakov[4], Teodor Nikolov[5], Zeyi Zhang[6], Tenglong Ao[6], Sicheng Yang[7], Shaoli Huang[8], Yongkang Cheng[8], M. Hamza Mughal[9], Rishabh Dabral[9], Kiran Chhatre[1], Christian Theobalt[9], Libin Liu[6], Stefan Kopp[2], Rachel McDonnell[10], Michael Neff[11], Taras Kucherenko[12], Youngwoo Yoon[13]*, Gustav Eje Henter[1,5]*

[1]KTH Royal Institute of Technology, [2]Bielefeld University, [3]University of Science – VNUHCM, [4]Independent Researcher, [5]Motorica AB, [6]Peking University, [7]Huawei Technologies Ltd., [8]Astribot, [9]Max-Planck Institute for Informatics, SIC, [10]Trinity College Dublin, [11]University of California, Davis, [12]SEED – Electronic Arts, [13]Electronics and Telecommunications Research Institute (ETRI)

Abstract

We review human evaluation practices in automated, speech-driven 3D gesture generation and find a lack of standardisation and frequent use of flawed experimental setups. This leads to a situation where it is impossible to know how different methods compare, or what the state of the art is.

In order to address common shortcomings of evaluation design, and to standardise future user studies in gesture-generation works, we introduce a detailed human evaluation protocol for the widely-used BEAT2 motion-capture dataset. Using this protocol, we conduct large-scale crowdsourced evaluation to rank six recent gesture-generation models – each trained by its original authors – across two key evaluation dimensions: motion realism and speech-gesture alignment.

Our results provide strong evidence that 1) newer models do not consistently outperform earlier approaches; 2) published claims of high motion realism or speech-gesture alignment may not hold up under rigorous evaluation; and 3) the field must adopt disentangled assessments of motion quality and multimodal alignment for accurate benchmarking in order to make progress.

Finally, in order to drive standardisation and enable new evaluation research, we will release five hours of synthetic motion from the benchmarked models; over 750 rendered video stimuli from the user studies – enabling new evaluations without model reimplementation required – alongside our open-source rendering script, and the 16,000 pairwise human preference votes collected for our benchmark.

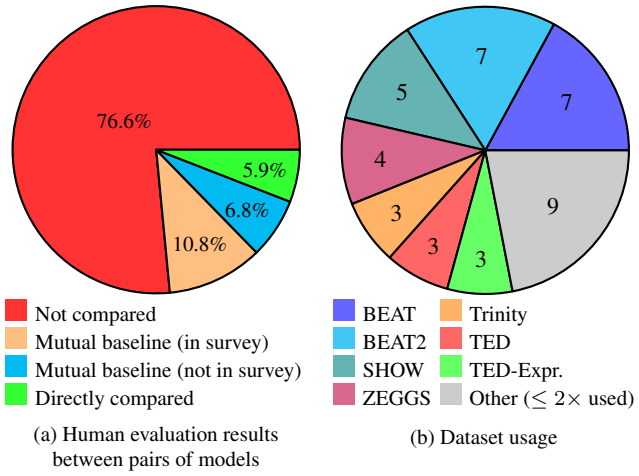
*Equal contribution.

1. Introduction

Research interest in automated gesture generation – the task of animating speaking 3D characters – has been sharply rising as part of the recent boom in generative and multimodal AI [1, 51, 59]. However, whilst the latest generative modelling techniques are being applied to this domain, trustworthy empirical evaluation of communicative non-verbal behaviour remains an understudied problem. Even human evaluations – widely considered the gold standard approach – may lead to misleading conclusions [12, 35] due to their complexity. Therefore, without a carefully designed evaluation standards adopted on a community level, published research may give a false sense of progress.

In this paper, we take a closer look at the evaluation practices of recent speech-driven gesture generation works, and identify several dimensions of evaluation setups where common design mistakes may lead to unreliable outcomes, and where the lack of standardisation makes similar studies incompatible. In order to address this problem, we develop an evaluation protocol on BEAT2 [47], the most widely adopted dataset according to our survey. While our protocol is rooted in the GENEA-Challenge methodology, we propose several improvements to increase the validity and re-usability of the evaluations.

To validate our BEAT2 evaluation protocol, and to understand the state of the art in gesture generation, we benchmark six recently published gesture-generation models using crowdsourced human evaluation following the proposed methodology. We then present a discussion of our findings, some of which contradict previously published claims, and release all collected information in the process, in order to enable future research on the evaluation of gesture generation. To summarise, we make three separate contributions



(a) Human evaluation results between pairs of models

(b) Dataset usage

Figure 1. Direct comparisons are exceedingly rare between state-of-the-art gesture generation models, partially due to spread out dataset usage. Data from 26 surveyed models published at CVPR, ICCV, ECCV, SIGGRAPH, and SIGGRAPH Asia between 2023 and 2025.

to address the lack of standardisation in gesture generation:

1. A critical **review** of evaluation practices (Sec. 3);
2. A new **evaluation protocol** for the widely-used BEAT2 dataset rooted in prior large-scale evaluations (Sec. 4);
3. A new **benchmark** of six recently published models to illuminate the state-of-the-art (Sec. 5).

2. Background

2.1. Automatic Gesture Generation

We define the task of speech-driven 3D gesture generation as the problem of mapping from a time-aligned input speech sequence s (e.g., an audio waveform and/or a text transcription) with potential additional conditioning c (e.g., an emotion label), to a 3D human body-motion sequence output x (e.g., a series of joint angles defining human poses). Whilst early deep-learning-based systems for gesture generation treated gesture generation as a regression problem [29, 33, 39–41, 71, 83, 84], recent approaches use probabilistic models like normalising flows [3], VAEs [30], VQ-VAEs [74, 81], discrete autoregressive transformer models [79], or diffusion models [4, 5, 20, 26, 55, 58, 78, 90].

2.2. Evaluations in Automatic Gesture Generation

Both automated ("objective") as well as human ("subjective") evaluation methodologies are widely used in 3D gesture generation [59]. While objective metrics like FGD [84] or beat consistency [50] can be useful to guide model development, they have well-known limitations and unclear relationships with human perception [24, 25, 32, 44, 73, 77, 79]. Therefore, final results are almost universally validated by analysing responses given by human evaluators.

> Human assessment must thus be considered the gold-standard evaluation methodology in gesture generation, and is the key focus of this paper.

2.2.1. Standardisation Needs

Despite being considered the gold standard, there is little information available on the ecological validity of human-evaluation practices in gesture generation. A 2021 review by Wolfert et al. [76] found a large diversity in both automated and human evaluation methods, with poor reporting practices on participant characteristics and evaluation design, concluding that the field would "*benefit from more experimental rigour and a shared methodology for conducting systematic evaluation*".

The GENEA Challenges [42–44, 85] were launched in 2020 to address this problem by evaluating sets of gesture-generation models under standardised conditions. The challenges are a series of community-driven evaluations of automated gesture-generation models. Each challenge collected between 5 and 12 model submissions trained on the same data, which were then subjected to large-scale crowdsourced human evaluation by the organisers. Challenge findings emphasise the importance of data filtering, removing artifacts, high-quality visualisations, and disentangled evaluation (as discussed in Sec. 3.1.1), and provide strong overall evidence for the importance of standardised human evaluation.

However, whether the GENEA Challenges address the problem of missing standardisation is an open question. There has been no assessment of whether their methodology is adopted by the community, and their results form isolated user studies rather than a continuously growing benchmark. Furthermore, the GENEA Challenges have seldom included submissions from major computer vision or machine-learning conferences like CVPR or SIGGRAPH, and therefore their results may not reflect the state of the art.

3. Key Limitations of Current Human-Evaluation Practices

Without standardised benchmarking practices, the state of the art can only be assessed by systematically reviewing independent evaluation results of published models. Are these evaluations reliable, or do inconsistencies in evaluation protocols, participant sampling, and reporting practices undermine their validity? Can we meaningfully assess the state of the art by comparing results from competing models when datasets, visualisation methods, and human-evaluation designs differ so widely across studies, or if such comparisons are fundamentally flawed?

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	Another model	Training dataset	Modelling Goal		Directly compared to...	
				Quality	Alignment	Mocap	Another model in the survey
2023	CVPR	DiffGesture [90]	TED, TED-Expressive	**Na**, **Sm**	**Rh**	✔	**B**
		QPGesture [79]	BEAT	**Hu**	**Gen**	✔	**B**
		RACER [70]	Trinity, own	**Re**	**Gen**, **Sem**	✗	
		TalkSHOW [82]	SHOW	✗	**Gen**	✔	**B**
	SG.	Bodyformer [60]	Trinity, TWH	**Hu**	**Gen**	✔	
		GestureDiffuCLIP [5]	BEAT, ZEGGS	**Hu**	**Sem**, **St**	✔	**B**
		LDA [4]	Trinity, ZEGGS	**Pref**	**St**	✔	**B**
	ICCV	C-DiffGAN [2]	PATS	**Na**	**Sem**, **Rh**, **St**	✔	
		LivelySpeaker [89]	BEAT, TED	**Na**, **Sm**	**Sem**	✗	
2024	CVPR	AMUSE [20]	BEAT2	✗	**Rh**, **Em**	✔	**M**
		Audio2Photoreal [58]	Audio2Photoreal	**Re**	✗	✔	**M**
		ConvoFusion [55]	DnD	**Na**	**Gen**, **Sem**	✔	
		DiffSHEG [16]	BEAT, SHOW	**Re**	**Rh**	✗	**M** **M** **B**
		EMAGE [47]	BEAT2	**Re**	✗	✗	**M** **B**
		EmoTransition [65]	BEAT-ETrans, TED-ETrans	**Na**, **Sm**	✗	✗	**M**
		ProbTalk [52]	SHOW	✗	✗	✗	**B**
	SG.	Sem. Gest. [87]	BEAT, ZEGGS, own	**Hu**	**Sem**, **Rh**	✔	**M**
	SG. Asia	SIGGesture [18]	BEAT	**Na**	**Sem**, **Rh**	✗	**M** **M** **M**
2025	CVPR	HOP [17]	TED, TED-Expressive	**Na**, **Sm**	**Sem**, **Rh**	✔	
		LOM [15]	BEAT2	✗	✗	✗	
		RAG-Gesture [56]	BEAT2	**Na**	**Sem**	✔	**M**
	SG.	MeCo [14]	BEAT2, ZEGGS	**Hu**	**Gen**	✗	**M**
	ICCV	GestureHydra [66]	SHOW, own	**Na**	**Sem**	✗	**M** **M**
		GestureLSM [49]	BEAT2	**Sm**, **Re**	**Rh**	✗	**M** **M** **M**
		SemGes [48]	BEAT, TED-Expressive	**Na**	**Sem**, **Rh**	✔	**M**
		SemTalk [86]	BEAT2, SHOW	**Re**	**Sem**, **Rh**	✗	**M**

Motivated by these questions, we perform a critical assessment of human evaluation practices in recent gesture-generation research, aiming to answer the following questions about recently published gesture-generation research:

1. Are evaluation results *reliable*, in that they measure what they purport to measure?
2. Are evaluation results *comparable* between different publications? Can we assess the state of the art from independent evaluations?

We review 26 recent publications on co-speech gesture-generation methods from selected computer vision and graphics conferences (CVPR, ICCV, ECCV, SIGGRAPH, and SIGGRAPH Asia) from 2023 onwards, using the search terms "gesture", "co-speech", "speech", and "motion" in publication titles, then filtering down the results to models whose outputs include 3D body gesture.

3.1. Entanglement Between Evaluation Dimensions

Upon analysing the human evaluation tasks, we find that gesture-generation research has converged to two high-level goals. In particular, every publication in Tab. 1 tries to assess some variation of the following:

1. *motion realism*, i.e., whether gestures look sufficiently natural and visually convincing to a human observer.
2. *multimodal alignment* between the output motion and the inputs (most commonly, speech).

Each of these aspects is crucial to applications of gesture generation, therefore it is important to assess them independently. A common method is to conduct two similar user

studies, one for each aspect, with only the evaluation question changed between the two. We will refer to this as the *naive setup*.

3.1.1. The Importance of Mismatching Evaluations

The first GENEA Challenge [42] used the naive setup to evaluate motion quality and multimodal alignment. Surprisingly, models with visually appealing motion received relatively high ratings for multimodal alignment even when they did not depend on the speech. In fact, Kucherenko et al. [44] re-analysed the data for all conditions in the GENEA Challenge 2020 and found a Pearson correlation of over 0.5 between the mean rated appropriateness of a given motion segment and its rated human-likeness. This is despite the fact that their crowdsourced test takers were instructed not to pay attention to human-likeness when assessing appropriateness. Therefore, there is strong evidence that:

> Direct evaluations of multimodal alignment – e.g., when conditioning on speech, emotion, style, or semantic information – may be ineffective unless the significant confounding effect of motion realism is isolated.

To combat the above problem, later GENEA Challenges [43, 44, 85] proposed a new evaluation paradigm leveraging the principle of *mismatching* [28, 37, 67, 85]. In this paradigm, so-called *mismatched stimuli* are created by swapping the data for a given modality to an unrelated sequence – for example, the audio is replaced by a different sentence, or the movements of the conversational partner are replaced by that of another character. Human evaluators, unaware of the mismatching manipulation, are then asked to indicate their preference between the original (matched) stimulus and its mismatched counterpart, based on which one seems more coherent (e.g., in terms of hand movement matching the speech). This setup effectively isolates the strong confounding effect of intrinsic motion quality from multimodal alignment, as the motion in both stimuli is always generated by the same system. Indeed, the GENEA Challenges 2022 and 2023 confirm the efficacy of this paradigm, reporting distinct model rankings for motion quality versus multimodal alignment [43, 85] and substantially reduced Pearson correlation between the two [44]. However:

> Our review finds that *none* of the multimodal-alignment evaluations in Tab. 1 adopt the mismatching methodology of the GENEA Challenges, nor do they employ other strategies to control for the confounding effect of motion quality.

This methodological gap is critical since it allows models that merely produce smooth, high-quality motion – regardless of how well they align with the speech content – to receive inflated multimodal alignment scores. Consequently, recent findings of speech-gesture alignment in model outputs closely approximating or surpassing the scores of human mocap (e.g., [56, 79, 82, 90]) are not reliable indicators of near-human performance. (We provide strong empirical support for this claim in Sec. 5.4.2).

3.2. Lack of Direct Comparisons

In this section, we investigate what kind of evaluation results are available when comparing the 26 models included in our survey. The strongest form of evidence for picking between two models comes from a direct comparison. If that is not available, it might still be possible to rank two models if one of them shows larger relative improvement compared to a mutual baseline model. However, if two publications consider non-overlapping sets of models in their evaluation, it becomes infeasible to infer which model works better without running a new evaluation.

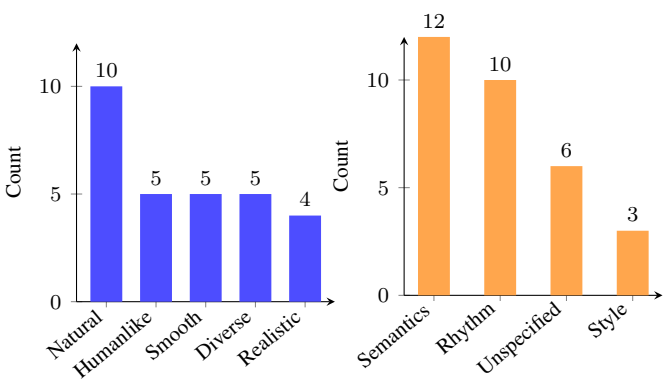
We uncover a remarkably low degree of direct comparisons between surveyed models, available only for 19 out of 325 model pairs. As visualised in Fig. 1, this amounts to **less than 6% of evaluation coverage of all possible pairs** of models in Tab. 1. Shockingly, almost three-quarters of the possible pairings between surveyed publications do not have relevant evaluation results! The right-hand pie chart highlights one reason for this fragmentation: the research community is divided across multiple datasets, with no single benchmark dominating usage. While BEAT and BEAT2 are the most common, a significant proportion of studies use smaller or proprietary datasets, which impedes reproducibility and consistent comparison even further.

It is of course unreasonable to expect every model to be compared to every other model in our review: baselines may be chosen from other conferences, and it is difficult to compare to models trained on other datasets or lacking publicly available implementation. However, human evaluations can depend strongly on which other systems are present in the evaluation, and the range of performance they exhibit [23]. For this reason, the lack of direct comparisons and relevant baseline choices between leading models, as uncovered by our review, pose a serious practical challenge for determining the state of the art.

3.3. Inconsistent, Incompatible Evaluation Designs

Design choices in, e.g., stimulus creation, question formulation, and test methodology can significantly affect the validity of human evaluation, and may introduce systematic biases. We find that current practices lack standardisation across these dimensions, and therefore it is generally infeasible to draw conclusions from cross-study comparisons.

Question Formulations We observe an ad hoc variability in how evaluations formulate the concept of motion real-



(a) Quality descriptors (b) Forms of alignment

Figure 2. Distribution of what aspects are measured in user studies evaluating motion realism (left) and multimodal alignment (right).

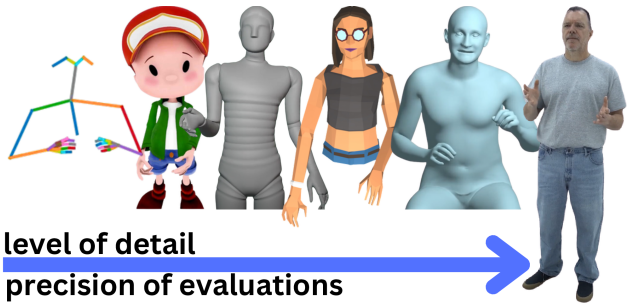


Figure 3. Example embodiments used in recent evaluations [20, 58, 60, 79, 85, 90], highlighting the high variety in 3D characters and their degree of realism. There is empirical support for the idea that more expressive visualisations make performance differences substantially easier to spot [58].

ism, with the usage of common adjectives shown on Fig. 2a. These terms lack a clear definition, and may encode different preferences – for example, "smooth" or "diverse" motion does not necessarily assume that the motion is "human-like" or "believable" While different adjectives could theoretically be used to evaluate distinct styles (e.g., "realism" may be more desirable for photorealistc avatars than for cartoon characters), our survey finds this is not the case in practice. Consequently, we believe that the inconsistency in question formulation needs to be addressed by the community, in order to avoid divergent and incomparable evaluation outcomes.

The lack of standardisation extends to evaluations of multimodal alignment (Fig. 2b). While the diversity of evaluation questions can reflect distinct modelling goals – such as semantic grounding, rhythmic alignment, and style control – we have already established that current evaluations setups may fail to isolate such nuances (cf. Sec. 3.1). We therefore postulate that these nominally specialised evaluations in practice measure the same underlying construct of general alignment. This motivates our proposal to establish a standardised, general-purpose alignment task as a foundational measure quantified through mismatching (Sec. 4).

Character Visualisation Character visualisation is another critical factor in human evaluation design [53]. Ideally, studies should use high-fidelity 3D avatars that can faithfully reproduce the original motion, avoiding artifacts from retargeting. Reinforcing this, a recent study by Ng et al. [58] found that realistic rendering, compared to untextured meshes, leads to substantially more distinguishing power in evaluations. Despite this, our survey finds that common practice falls short of this ideal. Several recent publications rely on simplistic stick-figure visualisations ([17, 65, 89, 90]), or untextured meshes ([4, 16, 20, 47, 82]). The majority of works in Tab. 1 employ unique, non-standardised 3D characters with substantial differences in realism, artistic style, and even perceived personality traits (Fig. 3). These differences introduce yet another confounding variable to evaluations, undermining the comparability and generalisation of study results.

Response Methodology The literature is further fragmented by the diversity of methodologies for presenting stimuli and gathering responses. Taking the widely-used BEAT and BEAT2 datasets as an example, we find evaluations collecting direct numerical ratings [65, 79], pairwise forced choice votes between models [47, 56], pairwise votes with ties [5, 20, 87], and ordinal ranking of more than two videos [16, 89]. The fundamental problem is not that these methods would produce different model rankings, but that their results are not mutually comparable. For instance, the mean rating of a model on a 5-point scale cannot be compared to the win-rate of another in a pairwise setup. Ultimately, the lack of a common voting protocol prevents meaningful cross-study comparisons, obscuring the community's understanding of the true state of the art.

3.4. Takeaways

Our analysis in Sec. 3 reveals three profound shortcomings of evaluation practices in gesture generation:

1. Doubtful ecological validity of evaluations of multimodal grounding due to the uncontrolled confounding factor of motion realism.
2. A systematic lack of direct comparisons between competing models.
3. No standardisation of methodological design and other implementation factors of evaluation setups.

Together, these three roadblocks make it impossible to reliably assess what the state of the art is, or understand which model is better for what purpose, without conducting the missing evaluations oneself. In fact, we theorise

that current evaluations may be more performative than informative (cf. Hertzmann [35]), possibly being more useful for paper acceptance than for actually measuring what they intend to measure.

4. Evaluation Protocol for the BEAT2 Dataset

Motivated by the findings of Sec. 3, we develop a human evaluation protocol for automatic 3D hand- and body gesture-generation models on the widely-used BEAT2 dataset. Our goals are to facilitate standardised evaluations, and to provide a template that can be easily extended to new datasets and modelling problems. Later, in Sec. 5, we validate our methodology by conducting a community-driven benchmarking of six competing models.

4.1. Evaluation Segment Selection and Rendering

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, imprecise results.

We curate 108 evaluation segments from the BEAT2 English test set, covering all speakers. The segments are randomly sampled complete sentences (according to the text transcription), 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.

To create the video stimuli shown to the crowdsourced evaluators, we develop a Blender-based rendering pipeline.To avoid retargeting errors, we directly visualise the SMPL-X models [61], with added textures for increased realism and the face hidden by a mask (as the unrelated facial motion may distract evaluators). For more details, please refer to Appendix A.8.

4.2. Evaluation Tasks

Following general practice (Sec. 3), we adopt the two evaluation tasks of motion realism and speech-gesture alignment. We construct a separate user study for each of the following evaluation questions:

1. Motion Realism: "In which video does the character gesture more like a real person?".
2. Speech-Gesture Alignment: "In which video do the character's movements fit the speech better?"

In contrast to common practice, we mute the audio during motion-realism evaluations. This important design choice allows us to evaluate the visual quality of synthetic gestures in complete isolation from how well they're aligned with the speech.

4.3. User-Study Methodology

In order to reduce the cognitive load of crowdsourced evaluators, we propose only using pairwise comparisons for evaluation. While potentially less time efficient than direct ratings (since each user response involves watching two video stimuli instead of just one), pairwise tests lead to higher inter-rater reliability [75], which is crucial given the inherent difficulty of evaluating co-speech gesturing.

We include five response options on Likert-type scale (weak- or strong preference in either direction, and a tie). To capture detailed user feedback in an economicalmanner, we adapt the *JUICE* methodology [31] – originally for video generation – to 3D motion evaluation for the first time. JUICE ("JUstify their choICE") requires test-takers to not only indicate their preferences, but also select predefined reasons for their choices.

In the next two sections, we describe the disentangled methodologies for our two evaluations. We refer the reader to the Appendix for implementation details such as precise question formulations, response options, JUICE factors, participant recruitment, attention checks, and more.

4.4. Motion-Realism Methodology

The wide range of ranking mechanisms in gesture generation evaluations prevent cross-study comparisons (Sec. 3.3). To solve this problem, we propose to use **Elo ratings** [11] under the Bradley-Terry model [11] as the standard ranking mechanism for motion realism in gesture generation. This methodology, originally for chess player ranking [27], has been successfully introduced to machine-learning evaluations by Chatbot Arena [88] (now LMArena).

A Bradley-Terry rating system assumes a fixed pairwise win-rate (under forced-choice comparison) for each pairing of gesture-generation models. Elo ratings are thereby computed using maximum-likelihood estimation such that the win-rate between two models A and B is a logistic sigmoid function of the difference between their respective Elo ratings R_A and R_B [11]. Following common practice, we use a base-10 logistic function with a scaling factor $S = 400$ such that:

$$P(\text{Model } A \text{ beats Model } B) = \frac{1}{1 + 10^{(R_B - R_A)/S}}, \quad (1)$$

Elo ratings offer several advantages. They are humanly interpretable: a difference of zero points means that the two systems are expected to beat each other half of the time (ignoring ties), whereas, e.g., a difference of +200 points (with our scaling factor) means that A is expected to be judged as

superior to B in about 76% of pairwise comparisons (averaged across all speech segments). Perhaps even more importantly, Elo ratings are inherently scalable as the number of evaluated models grows in a benchmark, since ranking between two models can be estimated from their Elo ratings even when direct comparisons are sparse or missing.

4.5. Speech-Gesture Alignment Methodology

Our literature review identified inaccurate multimodal alignment evaluation as a critical shortcoming (Sec. 3.1). The fundamental problem is that direct multimodal alignment ratings can be strongly confounded by the visual appeal of the motion, failing to distentangle the two aspects. We advocate for the mismatching methodology, introduced by the GENEA Challenges, as an effective solution.

Previous mismatching studies presented video clips with identical speech but different motion – one matched, one mismatched with the audio. However, Kucherenko et al. [44] found that differences in human-likeness ratings between individual motion segments (regardless of whether they were matched to the speech) explained a significant fraction of user preferences also in mismatching studies. To address this, we propose **audio mismatching**, where each video in a pair has the exact same motion, but different speech audio: one matched, and one mismatched from another evaluation segment. This distinguishes our evaluation from prior works [37, 43, 44, 67, 84, 85].

By keeping motion constant between compared video clips, the realism of the motion can no longer bias the evaluators. Unlike all prior gesture-generation evaluations, this setup therefore fully disentangles the evaluation of motion realism (where only the visuals differ between stimulus pairs) from the evaluation of speech-gesture appropriateness (where the visuals are identical between the two stimuli in a pair).

We note that there remains a potential bias between different speech segments, e.g., due to different voices being perceived as more appealing than others. We eliminate this bias by always using the same speaker voice in both stimuli. We further remove any systematic effect of possible preferences for individual speech segments within a speaker by ensuring that every speech segment appears equally often in matched as in mismatched videos.

5. Experiments

To address the gap in the community's understanding of the state of the art in gesture generation (Sec. 3), and to validate our human evaluation protocol (Sec. 4), we conduct a community-driven benchmarking effort. Due to known challenges with reproducibility in gesture generation, we invited authors of published models to participate by following our protocol, retraining their system under compatible settings, and submitting generated outputs for the full test set. In total, we compared six recently published gesture-generation models – DiffuseStyleGesture [78]; Semantic Gesticulator [87]; ConvoFusion [55]; RAG-Gesture [56]; AMUSE [20]; and HoloGest [19] – to each other and to the BEAT2 mocap recordings. For detailed information regarding the key features of these six systems, as well as a specification of how the retrained systems deviated from published models, please refer to Appendix B. We also experiment with JUICE factors and report those results in Appendix A.6 and Appendix A.7.

Additional results on automated metrics and their correlation with subjective human ratings are provided in Appendix C.

5.1. Motion-Realism Benchmark

We report the benchmark results obtained by following the motion realism protocol outlined in Sec. 4.4, and contrast them to prior results drawn from the status quo evaluation practices. We report Elo ratings with bootstrapped 95% confidence intervals; for statistical details, please refer to Appendix A.4.

5.2. Prior Results on Motion Realism

ConvoFusion and **RAG-Gesture** were evaluated with win rates in pairwise comparison, being preferred 43–47% percent of the time against human motion [55, 56]. **HoloGest** used a five-point Likert-type rating scale in an evaluation with 30 test takers and reported average human-likeness scores of 4.61 for natural motion, 4.47 for HoloGest, and 3.70 for DiffuseStyleGesture [19]. The user study in **Semantic Gesticulator** [87] did not find their system to be significantly different from human motion, whereas **AMUSE** [20] did. **DiffuseStyleGesture** received an average score of about 4.1 on a five-point scale, statistically indistinguishable from natural motion at 4.2 in their user study [78].

In summary, every system was originally reported to be close to the visual quality of human motion capture, as measured by user studies.

5.3. Our Results on Motion Realism

On Fig. 4, we report the Elo ratings acquired from our motion realism protocol (Sec. 4.4). The **Mocap** condition is at the top of the results with a mean Elo rating of 1133, setting an empirical upper bound for motion realism in this study. Amongst the machine-learning systems, **ConvoFusion** and **RAG-Gesture** rank as the strongest models with Elo ratings of 1102 and 1088, respectively. It makes sense for these systems to exhibit similar performance, given that RAG-Gesture builds on ConvoFusion. **HoloGest** follows closely behind RAG-Gesture with a mean Elo rating of 1084, slightly ahead of **Semantic Gesticulator** with an Elo

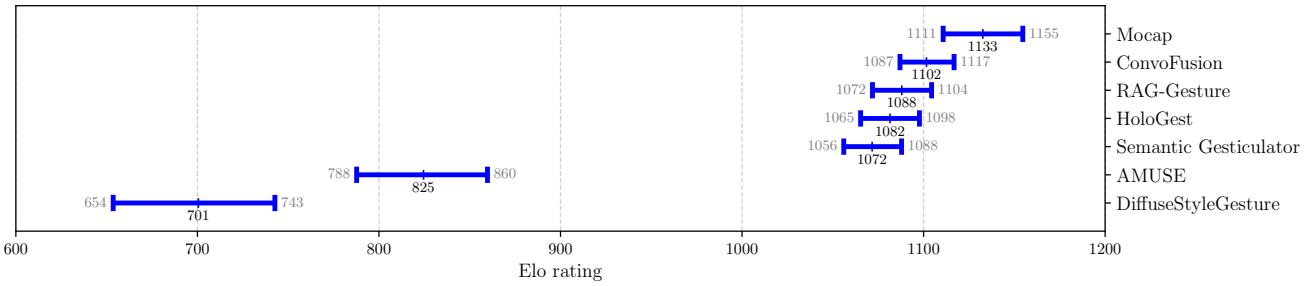


Figure 4. Results of the motion-realism user study, in the form of Elo ratings for each condition considered and 95% confidence intervals acquired from bootstrapping. Conditions are ordered by estimated Elo rating in descending order.

rating of 1070. All four of these systems remain well within a performance band that is arguably consistent with the broader state of the art.

At the lower end of the scale, **AMUSE** and **DiffuseStyleGesture** display notably lower Elo rating at 824 and 701, respectively. Notably, while AMUSE and DiffuseStyleGesture exhibit much wider confidence intervals, this does not imply that the comparative performance of these system is less predictable. Rather, this is a consequence of early stopping in our adaptive evaluation due to their clear separation from other conditions.

Our results found a considerable gap between the best-performing gesture-generation models and AMUSE and DiffuseStyleGesture, even though both those systems originally reported state-of-the-art performance. This may be taken as evidence that that previous claims of high motion quality may not hold up under careful evaluation, or that faithfully adapting published models to a new dataset or to new dataset splits can be a significant challenge.

More importantly, comparing the top systems, we find that:

> The BEAT2 dataset has become saturated for motion-realism evaluations, with four models showcasing comparable performance, with projected win rates between 41–46% against motion-capture recordings.

Importantly, this does not mean that the generated motion outputs are close to perfect. Rather, by intentionally removing the confounding factor of speech in our evaluations, we could show that the gap between gesture-generation models and natural human gesturing may be largely attributed to difficulties in aligning movements to the speech.

5.4. Speech-Gesture Alignment Benchmark

We also conduct the speech-gesture alignment evaluation as outlined in Sec. 4.5. We report bootstrapped 95% confidence intervals for audio mismatching scores, which are the weighted preference accuracy for the matched stimulus, independently computed for each condition, with strong preference votes counting twice. For more details, please refer to Appendix A.5.

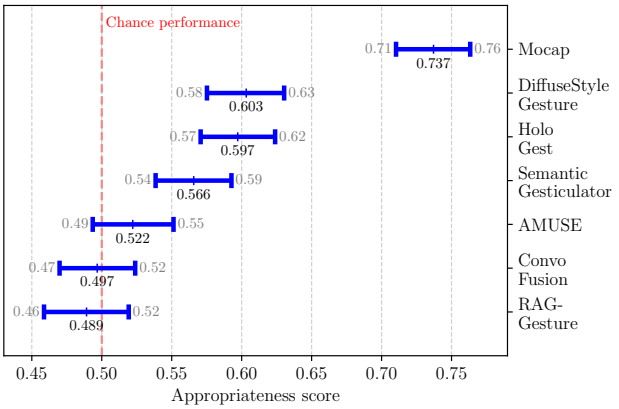


Figure 5. Results of the speech-gesture appropriateness user study. Ratio of user preference towards matched over mismatched stimuli for each condition considered, and 95% confidence intervals acquired from bootstrapping. Conditions are ordered by estimated appropriateness score in descending order.

5.4.1. Prior Results on Speech-Gesture Alignment

DiffuseStyleGesture performed a direct rating study which places its model at about an average score of 4.1 on a five-point scale, versus 4.2 for natural motion, calling it "competitive" with the latter [78]. The **HoloGest** paper again reports near-human mean opinion scores: 4.61 for human motion, 4.47 for their system, and DiffuseStyleGesture significantly behind at 3.91 [19]. **Semantic Gesticulator** reports a semantic accuracy score for BEAT data that is actually higher for their model (with a score of 0.41) than for human motion (with a score of 0.37), although the difference was not statistically significant [87]. **ConvoFusion** found a mean semantic-alignment score of 2.97 on a five-point scale, compared to 3.53 for natural motion [55], whilst **RAG-Gesture** states that its output was preferred over natural motion 44.6 percent of the time and proposes that it sets the state of the art [56]. Finally, **AMUSE** found

a gesture-synchronisation win/loss rate of 35%/52% compared to test-set motion, with 12% ties.

Five out of six systems reported multimodal alignment results nearing or exceeding that of human recordings.

5.4.2. Our Results on Speech-Gesture Alignment

As shown in Fig. 5, our speech-gesture alignment study exposes significant differences compared to prior published evaluations.

As expected, Mocap set the empirical topline with $\approx 74\%$ mean audio mismatching score. This far outperforms all generative models. DiffuseStyleGesture and HoloGest performed the best amongst the synthetic conditions, both scoring $\approx 60\%$. The near-identical alignment results suggest that HoloGest's previously reported alignment advances over DiffuseStyleGesture may in fact be confounded by motion quality, as postulated in Sec. 3.1.

Semantic Gesticulator reached $\approx 57\%$, sharply contrasting prior results indicating strong semantic alignment. AMUSE, ConvoFusion, and RAG-Gesture all have confidence intervals overlapping 50%, indicating that their speech-gesture alignment of is not statistically significantly different from chance rate without FDR correction. This can be taken as evidence that their output could be interchanged for random gesture sequences in this task (or at least for random gesture sequences aligned to the start of a sentence, since that is how our segments were selected).

Comparing these results to the motion-realism study highlights surprising contrasts. First, prior claims (apart from possibly AMUSE) greatly differ from our findings, with ConvoFusion and RAG-Gesture performing at chance level. Second, system rankings shifted substantially. DiffuseStyleGesture, the weakest system in realism, now matches the top generative performer, whilst ConvoFusion and RAG-Gesture exhibited no measured effect of speech-gesture alignment. The reversal suggests that design choices that optimise kinematic plausibility do not necessarily translate to good grounding in the speech, and that a model that pays explicit attention to speech-gesture alignment can outperform more advanced motion models on this performance measure.

The nearly identical scores of ConvoFusion and RAG-Gesture are unexpected, given that RAG-Gesture was designed to enhance semantic appropriateness via retrieval-augmented generation. However, this mechanism does not yet offer a measurable benefit in our current evaluation. This may be due to factors such as the retrieved samples not aligning sufficiently with the input speech or the synthesis stage introducing noise that masks any semantic gain. Another possibility is that the speech segments in our study may not provide enough opportunity for systems to demonstrate semantically grounded gesturing, as they were not pre-selected for this purpose.

Overall, we attribute the above differences between prior results and our findings to the standard evaluation practices failing to control for the confounding effect of motion realism, as discussed in Sec. 3.1, artificially inflating the measured appropriateness, given the high visual quality many of the evaluated systems have.

The use of mismatching, as featured here, is crucial in disentangling motion quality from appropriateness for speech. As it stands, the results in current literature are not reliable and convey a false sense of progress, where in reality we have very far left to go.

6. Limitations

Our conclusions for state-of-the-art performance – i.e., that motion realism is nearing saturation, while speech-gesture alignment is far from a solved problem – is naturally limited by the scale and quality of BEAT2, and may not extend to other datasets. Adapting our evaluation protocol to other datasets is an important future direction.

The use of Elo ratings [27] when evaluating motion realism comes with an assumption that the collective preferences of the test takers are transitive, which need not be the case; cf. Arrow [6]. Boubdir et al. [10] provides a discussion of this assumption in Elo-ratings-based evaluation of machine-learning models, along with possible alternatives. That said, the vast majority of existing gesture evaluations surveyed in Sec. 3 also assume transitivity, e.g., by having test takers assign ordinal ratings to different systems.

For speech-gesture alignment, we want to mention that the mismatching-based evaluation really measures *specificity* more than inherent appropriateness. If a gesture-generation system were to generate hypothetical "universally appropriate outputs" $x_{\text{universal}}$ that are a good fit to any possible speech audio segment s, such a system will score close to chance level in a mismatching-based appropriateness study. In practice, human motion contains rhythms at multiple levels [64], so a given motion x can be perceived as rhythmically appropriate to a range of different speech audio with different rhythms (cf. Avirgan [7], Miller et al. [54]), whereas semantic gestures (if produced by a system) are likely to be much more specific to a given context and speech. This may, among other things, cause our evaluations to be relatively more sensitive to semantic gestures and their appropriateness over rhythmic (beat) appropriateness. On the other hand, beats are vastly more common than semantic gestures, and are overall still likely to influence appropriateness evaluations more; cf. Saund and Marsella [69].

Our literature review only considers speech-driven 3D hand- and body gesture generation. There are several closely related tasks – e.g., solely text-driven generation,

facial motion synthesis, or video generation – that may suffer from similar problems, or offer alternative solutions to our methodology.

We also note that some of the systems in our evaluation required adaptations, and re-training, for the BEAT2 dataset. Although we mitigated these effects by having the original authors submit their outputs, such changes may nevertheless have an adverse effect on model performance, We hope that the authors of future models will perform first-party evaluation using our protocol and our released data, for more accurate measurement of the state-of-the-art.

Finally, the lack of reliable automated metrics is a significant problem. The human preference responses collected during our evaluation are suitable for training human opinion predictors [34], and may also enable the validation of existing and new automated metrics [24, 36].

7. Conclusion

In this paper, we performed a detailed survey of the various systematic problems in gesture-generation evaluations, declaring that it may not be possible to determine the state of the art from published results and that the lack of standardisation may contribute to a false sense of progress. We then designed a new evaluation protocol, based on best practices from prior evaluations, with added innovations (Elo ratings, audio mismatching, JUICE) for improved comparability and nuance without confounding. Finally, we put our protocol to the test by benchmarking six recent gesture-generation models, all of which have reported state-of-the-art performance.

Our evaluation results provide a first-of-its-kind benchmark between competing already published models, in a field severely lacking direct comparisons. We found strong evidence that motion realism on BEAT2 is no longer a distinctive factor between state-of-the-art models. Perhaps even more importantly, we show how current evaluation practices may lead to over-inflated multimodal-alignment results due to entanglement between evaluation dimensions. In contrast, our disentangled evaluations avoid common pitfalls, and pave the way towards more specialised evaluations for, e.g., semantic alignment or emotional expression.

Finally, we will release all collected data from our evaluations, including 16000 human votes, 750 rendered video segments, and five hours of synthetic motion from the benchmarked models. We hope that the proposed evaluation protocol and the released data will pave the way towards standardised gesture-generation evaluations [57] on currently used datasets, and enable the development of more targeted human evaluation methodologies and better automated metrics.

8. Acknowledgements

The authors thank Tu Anh Nguyen at FAIR for spotting a bug in an earlier version of our FGD evaluation code. RN, MT, TN, and GEH were partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. YY and GEH were partially supported by the Industrial Strategic Technology Development Program (grant no. 20023495) funded by MOTIE, Republic of Korea.

References

[1] Mohammad Mahdi Abootorabi, Omid Ghahroodi, Pardis Sadat Zahraei, Hossein Behzadasl, Alireza Mirrokni, Mobina Salimipanah, Arash Rasouli, Bahar Behzadipour, Sara Azarnoush, Benyamin Maleki, et al. Generative AI for character animation: A comprehensive survey of techniques, applications, and future directions. *arXiv preprint arXiv:2504.19056*, 2025. 1

[2] Chaitanya Ahuja, Pratik Joshi, Ryo Ishii, and Louis-Philippe Morency. Continual learning for personalized co-speech gesture generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20893–20903, 2023. 3

[3] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. Style-controllable speech-driven gesture synthesis using normalising flows. In *Computer Graphics Forum*, pages 487–496. Wiley Online Library, 2020. 2

[4] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–20, 2023. 2, 3, 5, 22

[5] Tenglong Ao, Zeyi Zhang, and Libin Liu. GestureDiffuCLIP: Gesture diffusion model with CLIP latents. *ACM Transactions on Graphics (TOG)*, 42(4):1–18, 2023. 2, 3, 5

[6] Kenneth J. Arrow. A difficulty in the concept of social welfare. *Journal of Political Economy*, 58(4):328–346, 1950. 9

[7] Jody Avirgan. Why spiderman is such a good dancer. https://web.archive.org/web/20201112011116/https://www.wnycstudios.org/podcasts/radiolab/articles/299399-why-spiderman-such-good-dancer, 2013. 9

[8] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 21

[9] Yoav Benjamini and Yosef Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stat.*, 25(1):60–83, 2000. 17

[10] Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. Elo uncovered: Robustness and best practices in language model evaluation. *Advances in Neural Information Processing Systems*, 37:106135–106161, 2024. 9

[11] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 6, 16

[12] Zoya Bylinskii, Laura Herman, Aaron Hertzmann, Stefanie Hutka, Yile Zhang, et al. Towards better user studies in computer graphics and vision. *Foundations and Trends® in Computer Graphics and Vision*, 15(3):201–252, 2023. 1

[13] Aggelina Chatziagapi, Louis-Philippe Morency, Hongyu Gong, Michael Zollhöfer, Dimitris Samaras, and Alexander Richard. AV-Flow: Transforming text to audio-visual human-like interactions. *arXiv preprint arXiv:2502.13133*, 2025. 19

[14] Bohong Chen, Yumeng Li, Youyi Zheng, Yao-Xiang Ding, and Kun Zhou. Motion-example-controlled co-speech gesture generation leveraging large language models. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, New York, NY, USA, 2025. Association for Computing Machinery. 3

[15] Changan Chen, Juze Zhang, Shrinidhi Kowshika Lakshmikanth, Yusu Fang, Ruizhi Shao, Gordon Wetzstein, Li Fei-Fei, and Ehsan Adeli. The language of motion: Unifying verbal and non-verbal language of 3d human motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3

[16] Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3, 5

[17] Hongye Cheng, Tianyu Wang, Guangsi Shi, Zexing Zhao, and Yanwei Fu. Hop: Heterogeneous topology-based multimodal entanglement for co-speech gesture generation. 2025. 3, 5

[18] Qingrong Cheng, Xu Li, and Xinghui Fu. Siggesture: Generalized co-speech gesture synthesis via semantic injection with large-scale pre-training diffusion models. In *SIGGRAPH Asia 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 3

[19] Yongkang Cheng and Shaoli Huang. HoloGest: Decoupled diffusion and motion priors for generating holisticly expressive co-speech gestures. In *Proceedings of the International Conference on 3D Vision*, 2025. 7, 8, 22

[20] 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*, pages 1942–1953, 2024. 2, 3, 5, 7, 19, 21

[21] Wei-Lin Chiang, Tim Li, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: New models & Elo system update. https://lmsys.org/blog/2023-12-07-leaderboard/, 2023. Accessed: 2025-05-20. 16

[22] Min Jin Chong and David Forsyth. Effectively unbiased FID and Inception score and where to find them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6070–6079, 2020. 22

[23] Erica Cooper and Junichi Yamagishi. Investigating range-equalizing bias in mean opinion score ratings of synthesized speech. In *Proc. Interspeech*, pages 1104–1108, 2023. 4

[24] Karlo Crnek, Grega Močnik, and Matej Rojc. Advancing objective evaluation of speech-driven gesture generation for embodied conversational agents. *International Journal of Human–Computer Interaction*, 0(0):1–17, 2025. 2, 10

[25] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[26] Anna Deichler, Shivam Mehta, Simon Alexanderson, and Jonas Beskow. Diffusion-based co-speech gesture generation using joint text and audio representation. In *Proceedings of the International Conference on Multimodal Interaction*, pages 755–762, 2023. 2

[27] Arpad E. Elo. The proposed USCF rating system, its development, theory, and applications. *Chess Life*, 22(8):242–247, 1967. 6, 9

[28] Cathy Ennis, Rachel McDonnell, and Carol O'Sullivan. Seeing is believing: body motion dominates in multisensory conversations. *ACM Transactions on Graphics (TOG)*, 29 (4):1–9, 2010. 4

[29] Ylva Ferstl and Rachel McDonnell. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents*, pages 93–98, 2018. 2

[30] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F Troje, and Marc-André Carbonneau. Zeroeggs: Zero-shot example-based gesture generation from speech. In *Computer Graphics Forum*, pages 206–216. Wiley Online Library, 2023. 2, 20

[31] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Factorizing text-to-video generation by explicit image conditioning. In *Proceedings of the European Conference on Computer Vision*, pages 205–224, 2024. 6

[32] Kazi Injamamul Haque, Alkiviadis Pavlou, and Zerrin Yumak. "wild west" of evaluating speech-driven 3d facial animation synthesis: A benchmark study. In *Computer Graphics Forum*, page e70073. Wiley Online Library, 2025. 2

[33] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. Evaluation of speech-to-gesture generation using bi-directional LSTM network. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents*, pages 79–86, New York, NY, USA, 2018. ACM. 2

[34] Zhiyuan He. Automatic quality assessment of speech-driven synthesized gestures. *International Journal of Computer Games Technology*, 2022, 2022. 10

[35] Aaron Hertzmann. The curse of performative user studies. *IEEE Computer Graphics and Applications*, 43(6):112–116, 2023. 1, 6

[36] Ali Ismail-Fawaz, Maxime Devanne, Stefano Berretti, Jonathan Weber, and Germain Forestier. Establishing a unified evaluation framework for human motion generation: A

comparative analysis of metrics. *Computer Vision and Image Understanding*, 254:104337, 2025. 10

[37] Patrik Jonell, Taras Kucherenko, Gustav Eje Henter, and Jonas Beskow. Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents*, 2020. 4, 7

[38] Maurice George Kendall. Rank correlation methods. 1948. 23

[39] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents*, pages 97–104, New York, NY, USA, 2019. ACM. 2

[40] Taras Kucherenko, Patrik Jonell, Sanne Van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 242–250, 2020.

[41] Taras Kucherenko, Dai Hasegawa, Naoshi Kaneko, Gustav Eje Henter, and Hedvig Kjellström. Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation. *International Journal of Human-Computer Interaction*, 37(14): 1300–1316, 2021. 2

[42] Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. A large, crowdsourced evaluation of gesture generation systems on common data: The genea challenge 2020. In *26th international conference on intelligent user interfaces*, pages 11–21, 2021. 2, 4, 15, 19

[43] Taras Kucherenko, Rajmund Nagy, Youngwoo Yoon, Jieyeon Woo, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. The GENEA Challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In *Proceedings of the International Conference on Multimodal Interaction*, pages 792–801, 2023. 4, 7

[44] Taras Kucherenko, Pieter Wolfert, Youngwoo Yoon, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. Evaluating gesture generation in a large-scale open challenge: The GENEA Challenge 2022. *ACM Transactions on Graphics (TOG)*, 2024. 2, 4, 7, 15, 19, 23

[45] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. AI choreographer: Music conditioned 3D dance generation with AIST++. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13401–13412, 2021. 22

[46] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. BEAT: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *Proceedings of the European Conference on Computer Vision*, pages 612–630, 2022. 22

[47] 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*, pages 1144–1154, 2024. 1, 3, 5, 21, 22

[48] Lanmiao Liu, Esam Ghaleb, Aslı Özyürek, and Zerrin Yumak. Semges: Semantics-aware co-speech gesture generation using semantic coherence and relevance learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 3

[49] Pinxin Liu, Luchuan Song, Junhua Huang, and Chenliang Xu. Gesturelsm: Latent shortcut based co-speech gesture generation with spatial-temporal modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 3

[50] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10462–10472, 2022. 2, 22

[51] Yu Liu, Gelareh Mohammadi, Yang Song, and Wafa Johal. Speech-based gesture generation for robots and embodied agents: A scoping review. In *Proceedings of the International Conference on Human-Agent Interaction*, pages 31–38, 2021. 1

[52] Yifei Liu, Qiong Cao, Yandong Wen, Huaiguang Jiang, and Changxing Ding. Towards variable and coordinated holistic co-speech motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1566–1576, 2024. 3

[53] Rachel McDonnell, Martin Breidt, and Heinrich H Bülthoff. Render me real? investigating the effect of render style on the perception of animated virtual humans. *ACM Transactions on Graphics (TOG)*, 31(4):1–11, 2012. 5

[54] Jared E. Miller, Laura A. Carlson, and J. Devin McAuley. When what you hear influences when you see: listening to an auditory rhythm influences the temporal allocation of visual attention. *Psychological Science*, 24(1):11–18, 2013. 9

[55] Muhammad Hamza Mughal, Rishabh Dabral, Ikhsanul Habibie, Lucia Donatelli, Marc Habermann, and Christian Theobalt. Convofusion: Multi-modal conversational diffusion for co-speech gesture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3, 7, 8, 21

[56] M. Hamza Mughal, Rishabh Dabral, Merel C. J. Scholman, Vera Demberg, and Christian Theobalt. Retrieving semantics from the deep: an rag solution for gesture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3, 4, 5, 7, 8, 19, 21

[57] Rajmund Nagy, Hendric Voss, Youngwoo Yoon, Taras Kucherenko, Teodor Nikolov, Thanh Hoang-Minh, Rachel McDonnell, Stefan Kopp, Michael Neff, and Gustav Eje Henter. Towards a genea leaderboard–an extended, living benchmark for evaluating and advancing conversational motion synthesis. *arXiv preprint arXiv:2410.06327*, 2024. 10

[58] Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard.

From audio to photoreal embodiment: Synthesizing humans in conversations. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3, 5, 19, 22

[59] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. A comprehensive review of data-driven co-speech gesture generation. In *Computer Graphics Forum*, pages 569–596. Wiley Online Library, 2023. 1, 2, 18, 19

[60] Kunkun Pang, Dafei Qin, Yingruo Fan, Julian Habekost, Takaaki Shiratori, Junichi Yamagishi, and Taku Komura. Bodyformer: Semantics-guided 3d body gesture synthesis with transformer. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. 3, 5, 19

[61] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 6

[62] 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 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 19, 20

[63] Olivier Perrotin, Brooke Stephenson, Silvain Gerber, and Gérard Bailly. The blizzard challenge 2023. In *18th Blizzard Challenge Workshop*, pages 1–27. ISCA, 2023. 15

[64] Wim Pouw, Shannon Proksch, Linda Drijvers, Marco Gamba, Judith Holler, Christopher Kello, Rebecca S. Schaefer, and Geraint A. Wiggins. Multilevel rhythms in multimodal communication. *P. Roy. Soc. B*, 376(1835), 2021. 9

[65] Xingqun Qi, Jiahao Pan, Peng Li, Ruibin Yuan, Xiaowei Chi, Mengfei Li, Wenhan Luo, Wei Xue, Shanghang Zhang, Qifeng Liu, and Yike Guo. Weakly-supervised emotion transition learning for diverse 3d co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10424–10434, 2024. 3, 5

[66] Kaisiyuan Wang Jiazhi Guan Shengyi He Fengguo Li Lingyun Yu Yingying Li Haocheng Feng Hang Zhou Hongtao Xie. Quanwei Yang, Luying Huang. Gesturehydra: Semantic co-speech gesture synthesis via hybrid modality diffusion transformer and cascaded-synchronized retrieval-augmented generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 3

[67] Manuel Rebol, Christian Güti, and Krzysztof Pietroszek. Passing a non-verbal Turing test: Evaluating gesture animations generated from speech. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces*, pages 573–581. IEEE, 2021. 4, 7

[68] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 22

[69] Carolyn Saund and Stacy Marsella. The importance of qualitative elements in subjective evaluation of semantic gestures. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021. 9, 19

[70] Mingyang Sun, Mengchen Zhao, Yaqing Hou, Minglei Li, Huang Xu, Songcen Xu, and Jianye Hao. Co-speech gesture synthesis by reinforcement learning with contrastive pre-trained rewards. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2331–2340, 2023. 3

[71] Kenta Takeuchi, Dai Hasegawa, Shinichi Shirakawa, Naoshi Kaneko, Hiroshi Sakuta, and Kazuhiko Sumi. Speech-to-gesture generation: A challenge in deep learning approach with bi-directional LSTM. In *Proceedings of the International Conference on Human Agent Interaction*, 2017. 2

[72] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 22

[73] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. EDGE: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023. 2

[74] Hendric Voß and Stefan Kopp. Aq-gt: a temporally aligned and quantized gru-transformer for co-speech gesture synthesis. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 60–69, 2023. 2

[75] Pieter Wolfert, Jeffrey M. Girard, Taras Kucherenko, and Tony Belpaeme. To rate or not to rate: Investigating evaluation methods for generated co-speech gestures. In *Proc. ICMI*, pages 494–502. ACM, 2021. 6

[76] Pieter Wolfert, Nicole Robinson, and Tony Belpaeme. A review of evaluation practices of gesture generation in embodied conversational agents. *IEEE Transactions on Human-Machine Systems*, 52(3):379–389, 2022. 2

[77] Karren D Yang, Anurag Ranjan, Jen-Hao Rick Chang, Raviteja Vemulapalli, and Oncel Tuzel. Probabilistic speech-driven 3d facial motion synthesis: new benchmarks methods and applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27294–27303, 2024. 2

[78] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. DiffuseStyleGesture: Stylized audio-driven co-speech gesture generation with diffusion models. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 5860–5868, 2023. 2, 7, 8, 20, 21

[79] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2321–2330, 2023. 2, 3, 4, 5

[80] Sicheng Yang, Haiwei Xue, Zhensong Zhang, Minglei Li, Zhiyong Wu, Xiaofei Wu, Songcen Xu, and Zonghong Dai. The diffusestylegesture+ entry to the genea challenge 2023. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 779–785, 2023. 20, 21

[81] Payam Jome Yazdian, Mo Chen, and Angelica Lim. Gesture2Vec: Clustering gestures using representation learning methods for co-speech gesture generation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3100–3107, 2022. 2

[82] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 469–480, 2023. 3, 4, 5

[83] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4303–4309. IEEE, 2019. 2

[84] 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 (TOG)*, 39 (6):1–16, 2020. 2, 7, 22

[85] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. The genea challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 736–747, 2022. 2, 4, 5, 7, 15, 19

[86] Xiangyue Zhang, Jianfang Li, Jiaxu Zhang, Ziqiang Dang, Jianqiang Ren, Liefeng Bo, and Zhigang Tu. Semtalk: Holistic co-speech motion generation with frame-level semantic emphasis. 2025. 3

[87] Zeyi Zhang, Tenglong Ao, Yuyao Zhang, Qingzhe Gao, Chuan Lin, Baoquan Chen, and Libin Liu. Semantic gesticulator: Semantics-aware co-speech gesture synthesis. *ACM Trans. Graph.*, 43(4), 2024. 3, 5, 7, 8, 19, 21

[88] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, pages 46595–46623, 2023. 6

[89] Yihao Zhi, Xiaodong Cun, Xuelin Chen, Xi Shen, Wen Guo, Shaoli Huang, and Shenghua Gao. Livelyspeaker: Towards semantic-aware co-speech gesture generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20807–20817, 2023. 3, 5

[90] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10544–10553, 2023. 2, 3, 4, 5

Gesture Generation (Still) Needs Improved Human Evaluation Practices: Insights from a Community-Driven State-of-the-Art Benchmark

Supplementary Material

A. User study set up and analysis

We now describe in more detail the experience of taking one of the user study, and how test takers were recruited.

A.1. Study participants

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. 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.

Crowdworkers who agree to take the test are first presented with an introduction screen presenting brief instructions about the task, with example screenshots of the response options. After that follow 25 test pages, each presenting a pair of video stimuli along with buttons and tick-boxes for test takers to provide their responses. An example of the test-taker GUI is presented in Fig. 7; the exact questions depend on the type of the user study, following Fig. 6.

A.2. Speech-Segment Selection

To particpiate in the evaluation, model authors were required to submit generated motion for the entire BEAT2 English test set, in long form (one long motion for each take in the test set). Submitters whose system is not deterministic, i.e., it may generate different output motion for the same inputs, are required to generate motion for the test set five times with different random seeds and submit all five generations, to allow deeper analysis of the diversity of their generated motion.

However, following prior large-scale evaluations in speech [63] and gestures [42–44], not all of the submitted motion was evaluated in the user-studies. Instead, a number of short *speech segments* – defined as start and end times in the test data – were selected, and only motion corresponding to these segments of speech are considered for the 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).

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.

Only keeping segments where the recorded motion is free of major artefacts is important, since our objective is to compare new gesture-generation methods against the realism of actual human motion, not against the realism achieved by the specific 3D motion-extraction methodology. (The presence of artefacts in the ostensibly natural motion in Kucherenko et al. [44], Yoon et al. [85] may explain their counter-intuitive finding that synthetic motion was preferred over natural motion for one of the systems they evaluated.) Concretely, we eliminated (1) all segments containing *flicking*, which we define as instant, physically impossible changes in pose, and most instances of (2) mesh penetration/self-intersection.

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 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.

A subselection of about 250 randomly chosen initial segments, drawn equally from all speakers, was manually whittled down to a a final set of 108 segments to be used in the evaluations on the BEAT2 dataset, only keeping segments whose motion had no major artefacts as described above. Motion-capture quality varied significantly across speakers, so we chose eight segments for speakers Wayne and Scott (which had the highest motion-capture quality) and four good segments for every other speaker in the test set. The chosen segments were between 7.7 and 12.0 seconds in duration (average 10.7 seconds) were selected. 108 segments is a significantly larger number of segments than used in previous large evaluations in the field, specifically the GENEA Challenges (which used 40–48 segments), yet still few enough that each stimulus will be assessed numerous times in the user studies, so as to allow future papers to perform stimulus-level statistical analyses of the data produced by our evaluation.

A.3. Attention Checks

Consistent with best practices in crowdsourced evaluations, we use 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 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).

A.4. Statistical Analysis for Motion Realism

Our statistical analysis is based on the pairwise preference data described above. We standardise the condition labels to canonical forms and convert the raw choices into triplets of the form *(model_A, model_B, 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 [21]. 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 [11], 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.

A.5. Statistical Analysis of Speech Appropriateness

For the statistical analysis, we use the basically same setup as described in Appendix A.4 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)

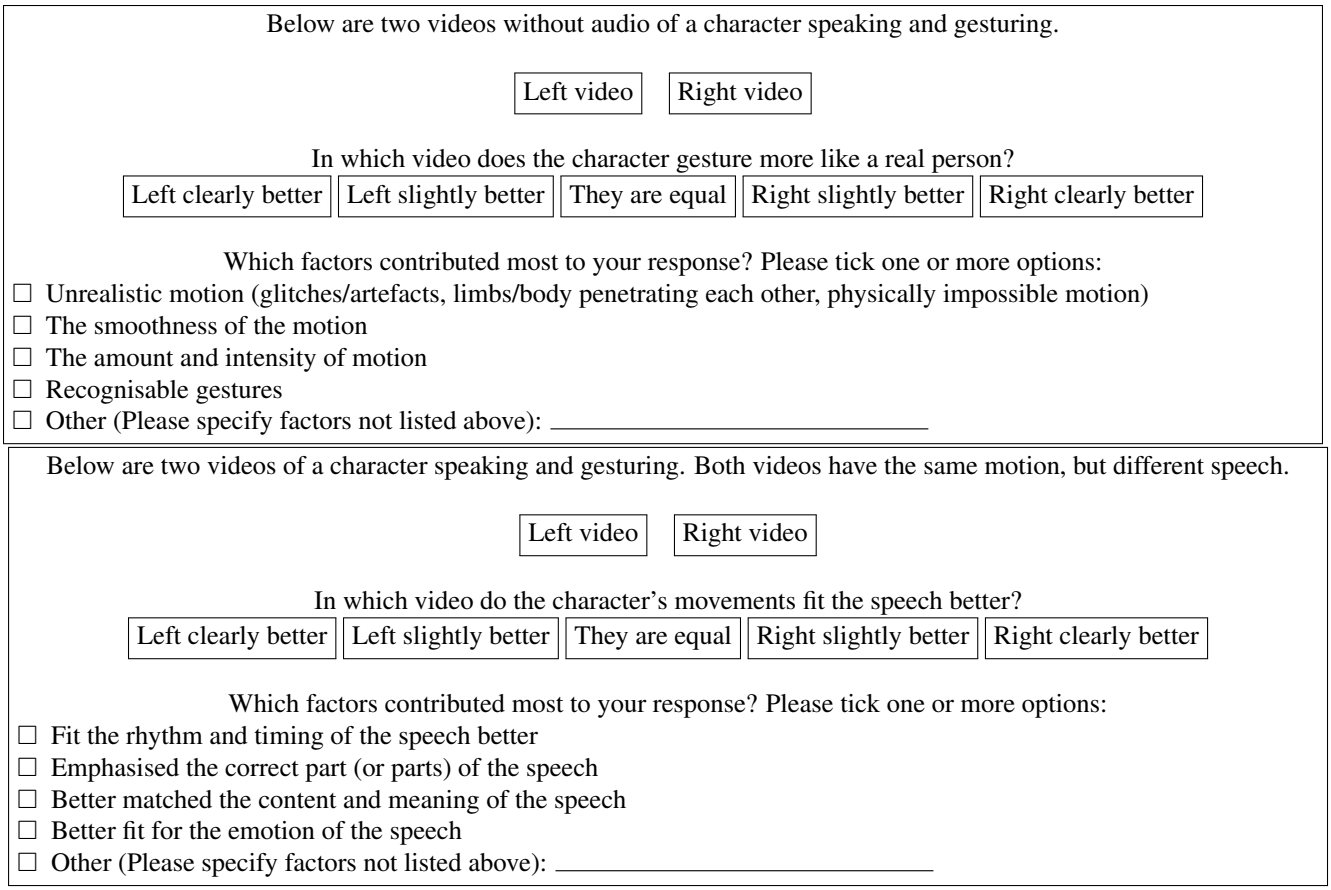


Figure 6. Questions and response options in the two types of user studies, also showing their schematic layout in the user-study GUI. For a screenshot of the GUI see Fig. 7.

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. To test if differences in appropriateness score between two conditions are statistically significantly different from zero, we likewise look at the differences between the appropriateness scores of any two systems in the bootstrap samples. We compute two-sided p-values regarding the score difference in the same way as for motion realism and apply Benjamini-Hochberg FDR correction [9] afterwards.

A.6. 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. 8. (All JUICE responses, including free text for the "Other" option, are featured in the data we release.)

Fig. 8 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 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.

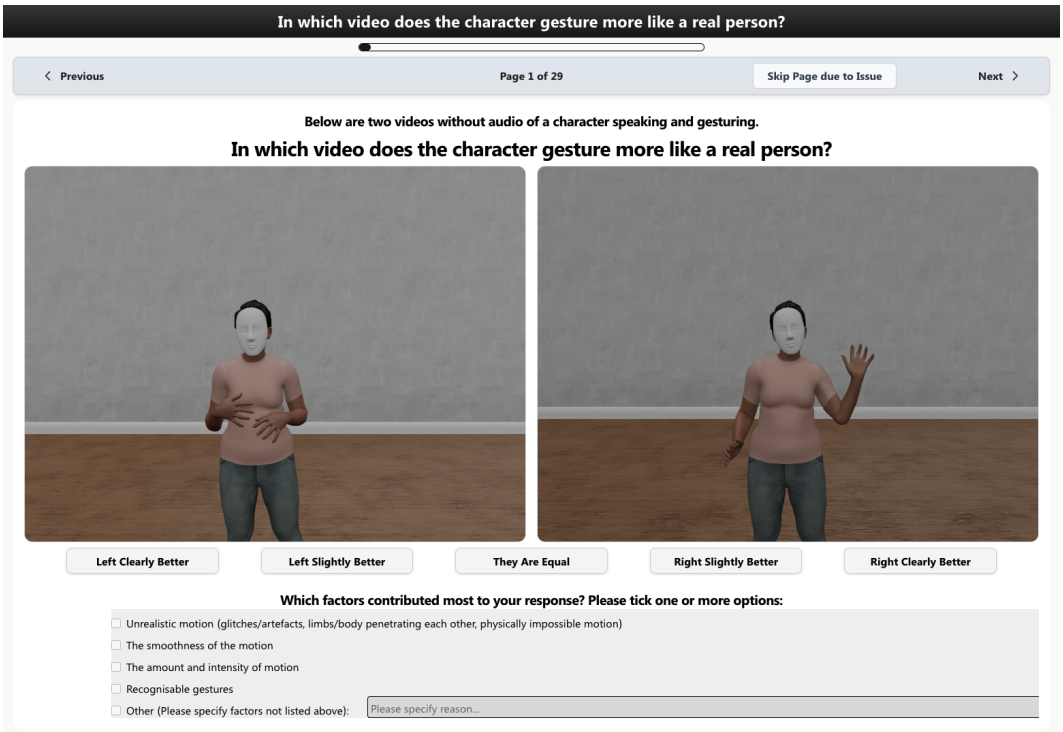


Figure 7. A screenshot of the GUI for the user studies, specifically from a motion-realism test with the current screen containing stimulus videos of the female avatar. The JUICE options are disabled with a grey background since no response to the main question has been selected yet.

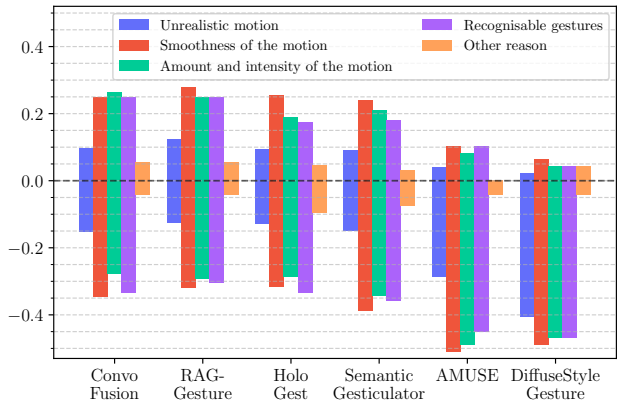


Figure 8. Frequency of JUICE options chosen for each model during the motion-realism evaluation in comparisons to the motion-capture condition, ignoring ties. Bar plots above zero show frequency among winning outcomes; bar plots below zero correspond to frequencies among losing outcomes, both relative to the total number of non-tie comparisons for the given model.

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 sur-

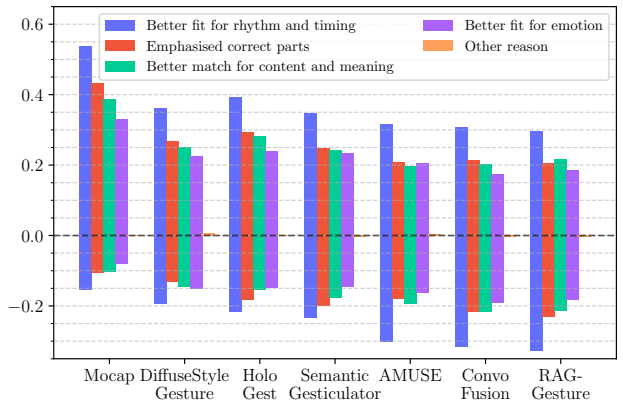


Figure 9. Frequency of JUICE responses chosen in the speech-gesture appropriateness study, for each model, when compared against its mismatched counterpart. Positive values are the frequency among winning outcomes; negative values correspond to the frequency among losing outcomes, both relative to the total number of non-tie comparisons.

prising 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 [56] 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 future that more clearly communicate its intention to crowdsourced test takers.

A.7. JUICE scores for gesture-speech appropriateness

The distribution of normalised JUICE responses for gesture–speech appropriateness is shown in Fig. 9. 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. 8 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. 9 but not in the earlier Fig. 8.

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 [69].)

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. 8, 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.

A.8. Gesture-Motion 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 [13, 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 in Kucherenko et al. [44], Yoon et al. [85], our visualisation (Fig. 10) 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 [20, 42, 60, 87].

The SMPL-X mesh offers an anatomically accurate body shape and proportion that matches each speaker [62]. 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

Figure 10. A video frame showing a gesturing SMPL-X avatar (male variant) rendered using the Blender visualiser.

does not replicate realistic 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 A.6). 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 C 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.

For consistency, all authors submit motion in SMPL-X format [62], and we render the video stimuli. This centralised rendering ensures uniform visual quality and prevents discrepancies that might otherwise arise from system-specific rendering pipelines.

B. 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.

B.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 prede-

fined 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 [47]. 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.

B.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.

B.3. ConvoFusion [55]

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 [8]. 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.

B.4. RAG-Gesture [56]

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 [55]), 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 B.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.

B.5. AMUSE [20]

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) [72] adapted to operate on filter-bank images. These vectors condition a latent-diffusion model [68] 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 A.2. 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.

B.6. HoloGest [19]

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 distribution 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.

C. 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 [47, 84]. **Fréchet Distance on Geometric and Kinetic Features (FD$_g$ and FD$_k$)** [4, 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 [45, 50]. **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 [46]. **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 2 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$_{pose}$. Semantic Gesticulator yields the best FGD and richest run-to-run variability (DIV$_{sample}$). Note that the FGD values are not directly comparable to those in Liu et al. [47] due to differing data sizes (see Chong and Forsyth [22]): 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

Table 2. 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}→	$\text{DIV}_{\text{sample}}$↑
	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.

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 appropriateness – showed the highest positive correlation (0.73). For speech-motion appropriateness, 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 appropriateness – underscore the limitations of current objective measures and the continued necessity of human evaluation.