

Same Motions, Different Winners: The Gesture-Generation Leaderboard and Its Human Agreement Are Configuration-Tunable

Deep Gandhi*
Transformer Lab

Ali Asaria
Transformer Lab

Tony Salomone
Transformer Lab

Abstract

Fréchet Gesture Distance (FGD) is the standard automatic metric for co-speech gesture generation, the models that synthesize the hand and body motion accompanying speech. Human perceptual studies remain the gold standard, and the GENE challenge ranks systems by them, but such studies are slow and costly, so papers routinely compare systems by FGD instead. FGD is a Fréchet distance in the latent space of a learned pose autoencoder, so it inherits the many choices behind that encoder, most of which papers never report. We ask whether FGD is a reliable proxy for perceptual evaluation under these unreported choices. Holding the systems fixed, we vary the reference tool’s own settings (analysis window, seed, feature versus raw space) and the corpus the feature extractor is trained on, on both tiers of the GENE 2022 evaluation and on the GENE 2023 challenge. First, the FGD ranking is not reproducible where it matters: the “best” system is one of two that a per-clip bootstrap shows are statistically tied, so which one wins is a configuration choice, though aggregate rank correlation stays moderate rather than collapsing. Second, and more consequential, FGD’s agreement with human ratings is modest and itself tunable, moving by 0.18 to 0.24 Kendall τ across these undocumented choices: FGD can be made to look appreciably more or less human-valid without changing the systems. These effects replicate across two challenge years and three evaluation tiers. We conclude that FGD numbers are not comparable across papers unless the window, seed, scoring space, and feature-extractor training corpus are pinned and reported, and that FGD cannot resolve small differences between closely performing systems.

1 Introduction

Co-speech gesture generation synthesizes the hand and body motion that accompanies speech, from

audio and text, to animate embodied conversational agents and virtual avatars (Nyatsanga et al., 2023). Human perceptual studies are the gold standard for evaluating such motion, and the GENE challenge ranks systems by large crowdsourced user studies (Yoon et al., 2022; Kucherenko et al., 2024, 2023). Because such studies are slow and costly, papers routinely compare systems with Fréchet Gesture Distance (Yoon et al., 2020) instead, the automatic metric the challenges report alongside their human evaluation and the one surveys treat as standard (Nyatsanga et al., 2023; Nagy et al., 2024). FGD is the gesture analogue of FID (Heusel et al., 2017) and FVD (Unterthiner et al., 2018): a Fréchet distance in the latent space of a pose autoencoder trained to reconstruct motion. Because it is a distance in a *learned* feature space, it inherits every choice made in building that space, such as the corpus the encoder saw, the motion window length, the pose representation, and the random seed. These choices are rarely reported, yet FGD numbers are compared across papers that made different choices.

We ask whether FGD can be a reliable proxy for perceptual evaluation when only these nuisance choices change. Holding the systems fixed, we vary two families of configuration on the GENE 2022 challenge outputs and their released human-likeness ratings: (A) the reference tool’s exposed settings (window length, seed, and feature versus raw scoring), and (B) the corpus the feature extractor is trained on. We measure the stability of the induced ranking and its agreement with the human ranking, and repeat everything on the second GENE 2022 tier and on the independent GENE 2023 challenge.

We find two things. The FGD ranking is not reproducible at the decision-relevant margin, though aggregate correlation stays moderate rather than collapsing. More consequentially, FGD’s agreement with human judgment is modest and can be

*Corresponding author: deep@lab.cloud

moved appreciably by these undocumented choices. Because the choices are neither standardized nor reported, two papers computing FGD with different but reasonable settings can report different winners and different apparent validity without any change to the systems: FGD is unreliable for resolving small differences between closely performing systems, and cross-paper FGD numbers are not comparable. Our contributions are (i) a controlled audit of FGD across two GENE challenge years that varies both the tool’s exposed settings and the feature-extractor training corpus, measuring ranking stability and human agreement together; (ii) evidence that the top system flips between statistically tied systems while aggregate correlation stays moderate; and (iii) the main result, that FGD’s human agreement is a researcher-tunable quantity, with concrete reporting recommendations.

2 Related Work

Fragility of Fréchet-style metrics. Fréchet-style generative metrics are known to be fragile to exactly the design choices we vary. Studies of FID, FVD, and Fréchet Audio Distance document content bias (Ge et al., 2024), sensitivity to sample size and the Gaussian assumption (Bischoff et al., 2024; Bench et al., 2026), misranking relative to human or downstream judgment (Wu et al., 2025; Jiralerspong et al., 2023), and dependence on the feature extractor and the choice of two-sample test (Chung et al., 2025; Cai et al., 2025; Grossi et al., 2025). These establish that the fragility exists in principle; what is missing is a measurement of how far it moves a real set of rankings.

Evaluating gesture generation. In gesture generation, human perceptual studies remain the gold standard (Kucherenko et al., 2024, 2023), but they are slow and costly, which is why FGD is used as their automatic proxy. Several works question that proxy: the GENE Leaderboard position paper (Nagy et al., 2024) argues qualitatively for a standardized living benchmark, “2D or not 2D” (Guichoux et al., 2024) shows pose-representation dimensionality shifts FGD values for a pair of systems, and gesture perception has been examined in immersive settings (Werner et al., 2025). Closest to our setting, Crnek et al. (2025) search 56 feature-extractor and distance combinations on GENE 2022 and 2023 for ones that better correlate with human studies (best ≈ 0.81); they seek a better metric, whereas we hold the standard FGD fixed

and measure how far its undocumented configuration moves the rankings that papers already report, across two challenge years.

Robustness of automatic leaderboards. Our framing connects to a growing concern with the robustness of automatic leaderboards, from benchmarks that prove fragile to superficial changes (Feng et al., 2025) to leaderboards whose integrity can be undermined through the evaluation pipeline (Naseh et al., 2025). FGD-based rankings carry a similar but entirely undocumented degree of freedom in the metric’s own configuration. Prior work asks which extractor best correlates with human studies; to our knowledge we are the first to audit the *ranking-comparability* consequences of the standard FGD’s undocumented configuration, measuring the induced ranking instability and the change in human agreement together.

3 Method and Setup

Metric. FGD fits Gaussians to the pooled feature vectors of a candidate system’s motion and of natural motion, and returns their Fréchet distance (Yoon et al., 2020); the feature space is the latent of an EmbeddingNet pose autoencoder. We use the GENE reference implementation (Kucherenko et al., 2024),¹ train the autoencoder for 10 epochs (Adam, learning rate $5e-4$; the released code uses 50, which we also test), and score every system against natural motion. Our default extractor is trained on the pooled motion of all systems, natural included. The challenge paper describes its extractor as “trained on the challenge dataset”; the released code trains it on the natural-motion test clips. Axis B below includes that natural-only corpus, so both readings are covered. The unit is a *system* (one model’s released outputs); we never retrain a generator.

What we vary. The manipulated variable is the FGD configuration, along two axes. **Axis A** (the tool’s settings) is a grid over window length {30, 60, 90} frames, seed {0, 1}, and scoring space (the 32-dimensional latent, “feat”, versus mean-pooled raw pose, “raw”), giving 12 configurations. **Axis B** (the training corpus) holds window at 30 and feature space, and trains the autoencoder on five corpora drawn from the same pool: all systems, natural only, all synthetic, and two disjoint halves

¹github.com/gene-workshop/gene_numerical_evaluations

of the systems. The community treats the extractor as neutral; this axis tests that directly.

What we measure. We rank systems by FGD and compare rankings with Kendall’s τ_b : the number of system pairs ordered the same way in both rankings minus the number ordered differently, divided by the number of pairs, with a correction for ties. $\tau = 1$ means identical order, 0 no association. With 10 synthetic systems there are 45 pairs, so τ moves in steps of $2/45 \approx 0.044$, and $\tau = 0.867$ means three pairs are ordered differently. We report three quantities. *Stability* is τ between the rankings of two configurations, plus whether the top-ranked system changes (a top-1 flip). *Validity* is τ between a configuration’s FGD ranking and the human ranking, where the human ranking orders systems by their median human-likeness rating. *Spread* is the range of validity across configurations. We target human-likeness because FGD is a motion-realism metric matched to that axis, whereas the challenges’ appropriateness rating turns on speech-gesture alignment that FGD does not model. Natural motion is the reference, so it attains near-zero FGD and trivially wins; headline numbers therefore use the *synthetic-only* ranking (natural excluded), with the with-natural counterpart reported alongside. Feat and raw are different metrics, so feat-only stability is primary. We quantify uncertainty by bootstrapping over study participants (validity), over the 40 clips per system (the FGD values), and by retraining at a larger budget (to rule out under-training).

Data. We use the GENE Challenge 2022 (Yoon et al., 2022), whose evaluation is analysed in extended form by Kucherenko et al. (2024):² the upper-body tier has 11 systems (10 synthetic plus natural), 40 clips each, pose dimension 162, 26,785 windows at length 30; the full-body tier has 10 systems and pose dimension 174, used for replication. We add a second challenge year, GENE 2023 (Kucherenko et al., 2023)³ (15 systems including natural, 30 clips each), whose outputs are released only as joint-angle BVH, so we convert them to 3D joint coordinates (root-relative, 330 dimensions) with the challenge’s own forward-kinematics tooling, verified by a near-zero natural-versus-natural FGD. Systems carry anonymized GENE identi-

fiers (for example USQ, USM). Ratings are the released per-participant scores (roughly 1000 to 1500 per system, upper-body). All motion and ratings are public Zenodo releases (footnotes 2 and 3); no generator was trained.

4 Results

Ranking stability (Axis A): the aggregate order is moderately stable, but the winner is not.

Across the 12 upper-body configurations, synthetic-only pairwise τ has median 0.867 and minimum 0.733; within the feature space alone it is higher (0.911 / 0.822). Rankings are never identical but the aggregate correlation is moderate, not a collapse: a typical configuration pair inverts a median of 3 of the 45 synthetic pairs, the most discordant pair inverts 6. Including natural raises stability slightly (0.891 / 0.782). What breaks is the top of the table: USQ ranks first in 4 of 12 configurations and USM in the other 8, so two systems are each “best” depending only on window, seed, and scoring space, and the flip occurs within the feature space. At window 30 the winner flips with the seed alone, and this persists when the extractor is retrained at five times the budget (50 epochs), so it is not an undertraining artifact. A per-clip bootstrap (500 resamples over the 40 clips; window 30, feature space, seed 0) shows USQ and USM are *statistically tied*: FGD 3.09 (95% CI [2.21, 4.56]) and 3.70 (95% CI [2.51, 5.91]), gap CI [−1.35, 3.11], positive in 71% of resamples. The top-1 flip is thus a tie-flip: FGD cannot separate the two systems yet assigns a definite rank that flips with configuration, which is the mechanism behind our recommendation to treat close differences as ties.

Validity: human agreement is modest and moves with configuration.

We treat validity as exploratory. Against the human ranking, synthetic-only τ has median 0.422, so FGD captures under half of the pairwise order of human preference. The disagreement is specific: at a median-agreement configuration (window 60, seed 0, feature space) FGD’s best synthetic system USM is humans’ 7th of 11, FGD’s worst system USN is humans’ 5th, and the human favourite USQ is FGD’s 3rd. Including the natural anchor (which FGD trivially places first) raises the median to 0.491, matching the $\tau \approx 0.5$ FGD–human agreement reported by Kucherenko et al. (2024); the lower synthetic-only figure is the harder natural-excluded test, not a discrepancy with prior work. This agreement is tun-

²Motion: zenodo.org/records/6973297; ratings: zenodo.org/records/6940057.

³Motion: zenodo.org/records/8146027; ratings: zenodo.org/records/8434117.

able: across the 12 configurations it ranges from 0.333 to 0.511 (spread 0.178, standard deviation 0.057; Table 1), which corresponds to roughly four of the 45 system pairs reordering. Window length is one axis, but its direction is confounded: shorter windows yield more windows (window 30 about three times window 90) and Fréchet distance is sample-size biased (Bischoff et al., 2024). Holding the per-system window count fixed, the apparent shorter-is-better trend disappears and reverses (matched τ 0.289, 0.467, 0.333 at windows 30/60/90), so we make no recommendation on window length; the point is that agreement moves with the window. Window 30 and window 90 at their natural window counts (seed 0, feature space) do differ under participant resampling (paired difference 0.133, 0 of 2000 resamples ≤ 0 , one-sided $p < 0.0005$; the marginal CIs overlap because they retain between-participant variance that the paired difference cancels, and τ over ten systems is quantized in 0.044 steps). Since those two configurations also differ in window count, this shows the choice is consequential, not that shorter is better.

Axis B: the extractor’s training corpus is a hidden axis. Training the autoencoder on five corpora from the same pool yields synthetic-only cross-corpus τ of median 0.800 and minimum 0.644, comparable to, if not larger than, the reordering from the exposed settings. The mid-table moves (USN shifts from rank 10 to 7 by corpus), though the top is stable on this tier. Per-corpus human agreement ranges from 0.378 (all, natural, or all synthetic) to 0.556 (one arbitrary half), standard deviation 0.069: an undocumented, unmotivated choice moves apparent validity. A size-matched control (every corpus subsampled to the same window count before training the extractor) reproduces the reordering (synthetic-only τ 0.778 / 0.600) and most of the human-agreement spread (0.133 of 0.178), so the axis is primarily corpus content, not sample size. One confound remains: every corpus is drawn from the systems being scored, so we read this result as showing that the extractor is not neutral, without claiming which corpus is the right one.

Replication (full-body and GENE 2023). Both effects reproduce on the full-body tier (Table 1). Axis A synthetic-only τ is 0.889 / 0.722, and the top system flips between FSA and FSH. Axis B cross-corpus τ is 0.806 / 0.667, and here the instability reaches the top: training on an arbi-

Axis / tier	Stability τ		Validity τ	
	median	min	range	spread
A: settings (upper)	0.867	0.733	0.333–0.511	0.178
A: settings (full)	0.889	0.722	0.333–0.556	0.222
A: settings (2023)	0.802	0.626	0.451–0.692	0.242
B: corpus (upper)	0.800	0.644	0.378–0.556	0.178
B: corpus (full)	0.806	0.667	0.389–0.611	0.222

Table 1: Stability (synthetic-only pairwise Kendall τ between the FGD rankings of different configurations) and validity (synthetic-only Kendall τ between each configuration’s FGD ranking and the human ranking, as a range over configurations and its spread). Rankings are never identical, the training corpus reorders the board comparably to the tool’s own settings, and human agreement is modest and moves with configuration.

trary half flips the winner from FSA to FSH. Human agreement is again modest and tunable, with spread 0.222 on both axes. Which corpus agrees best differs by tier (full-body favours natural-only at $\tau = 0.611$), so the invariant is the large spread, not a consistent best corpus, which is itself evidence the choice is an unprincipled degree of freedom. The finding also holds on the GENE 2023 challenge, a different year with different systems and a different skeleton: synthetic-only τ is 0.802 / 0.626 and the top synthetic system again flips (SJ versus SG), while agreement with humans is again modest (median 0.571) and configuration-dependent (spread 0.242).

5 Discussion

No single FGD number is comparable across papers unless the window, seed, scoring space, and feature-extractor training corpus are all pinned and reported: two papers making different but reasonable choices can disagree on which system is best and on how human-valid FGD appears, without any difference in the systems. We recommend that FGD be reported with a fully disclosed configuration including the training corpus; that small FGD differences not be interpreted, since FGD cannot separate closely performing systems and even the top two here are statistically indistinguishable under clip resampling; and that cross-paper FGD comparisons be treated as unsound unless these choices are held identical.

6 Conclusion

FGD rankings of gesture-generation systems, and FGD’s very agreement with human judgment, are

tunable by configuration choices that papers do not report. The ranking is not reproducible at the decision-relevant margin, where even the top two systems are statistically tied, and the metric’s apparent validity can be moved by as much as 0.18 to 0.24 Kendall τ across tiers without changing the systems. These effects hold across two GENEA challenge years and three tiers. Comparable gesture evaluation requires pinning and reporting the full FGD configuration, including the feature-extractor training corpus, and treating close FGD differences as ties.

Limitations

The evidence is on the GENEA challenge datasets (two challenge years, three tiers); the direction should generalize to other corpora and motion formats, though magnitudes may not. Aggregate correlation is moderate (0.80 to 0.89), so the honest reading is that FGD is unreliable at the margin and for close systems, not that it is random. The training-corpus axis uses in-format subsets of the scored pool rather than an external corpus (the challenge training data is in a different motion format); a size-matched control rules out sample size, but the extractor still trains on the systems it scores, so a fully disjoint external corpus remains future work. The window-length trend does not survive matching sample size, so we claim no window-length effect; and the participant bootstrap covers rating noise, not FGD-side noise except where we bootstrap the FGD values. Kendall τ_b over ten systems is coarse (steps of 2/45), which we reflect by reporting worst-case and typical discordance separately.

Availability

The code, runner, reference outputs, and data recipe are publicly available.⁴ The audit uses only the public GENEA motion and rating releases cited in Section 3, which we do not redistribute.

Acknowledgements

The experiment setup and planning for this study were carried out with an earlier version of Primus, an automated research assistant developed by Transformer Lab.

⁴github.com/transformerlab-research/fgd-config-audit

Generative AI Usage

The experiment setup and planning used Primus, an LLM-based research assistant (see Acknowledgements). LLM tools (Claude Opus 4.8) were also used to format tables and to improve language clarity and grammar. The authors verified every reported number against the run outputs and take full responsibility for the content.

References

- Ciaran Bench, Vivek Desai, Carlijn Roozemon, Ruben van Engen, and Spencer A. Thomas. 2026. [Trustworthiness of Fréchet Inception Distance via Monte Carlo Dropout](#). *arXiv preprint arXiv:2601.21979*.
- Sebastian Bischoff, Alana Darcher, Michael Deistler, Richard Gao, Franziska Gerken, Manuel Glöckler, Lisa Haxel, Jaivardhan Kapoor, Janne K. Lappalainen, Jakob H. Macke, Guy Moss, Matthijs Pals, Felix Pei, Rachel Rapp, A. Erdem Sagtekin, Cornelius Schröder, Auguste Schulz, Zinovia Stefanidi, Shoji Toyota, and 2 others. 2024. [A Practical Guide to Sample-based Statistical Distances for Evaluating Generative Models in Science](#). *Transactions on Machine Learning Research*, 2024. ArXiv:2403.12636.
- Yuting Cai, Shaohuai Liu, Chao Tian, and Le Xie. 2025. [Fréchet Power-Scenario Distance \(FPD\) for Smart Grids](#). *arXiv preprint arXiv:2505.08082*.
- Yoonjin Chung, Pilsun Eu, Junwon Lee, Keunwoo Choi, Juhan Nam, and Ben Sangbae Chon. 2025. [KAD: No More FAD! Kernel Audio Distance for Audio Generation](#). *arXiv preprint arXiv:2502.15602*.
- Karlo Crnek, Grega Močnik, and Matej Rojc. 2025. [Advancing Objective Evaluation of Speech-Driven Gesture Generation for Embodied Conversational Agents](#). *International Journal of Human-Computer Interaction*. Published online July 2025.
- Haiwen Feng, Long Lian, Lisa Dunlap, Jiahao Shu, XuDong Wang, Renhao Wang, Trevor Darrell, Alane Suhr, and Angjoo Kanazawa. 2025. [Visually Prompted Benchmarks Are Surprisingly Fragile](#). *arXiv preprint arXiv:2512.17875*.
- Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. 2024. [On the Content Bias in Fréchet Video Distance](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7277–7288. IEEE. ArXiv:2404.12391.
- Samuele Grossi, Marco Letizia, and Riccardo Torre. 2025. [Refereeing the Referees: Evaluating Two-Sample Tests for Validating Generators in Precision Sciences](#). *Machine Learning: Science and Technology*, 6(1):015052. ArXiv:2409.16336.

- Téo Guichoux, Laure Soulier, Nicolas Obin, and Catherine Pelachaud. 2024. [2D or not 2D: How Does the Dimensionality of Gesture Representation Affect 3D Co-Speech Gesture Generation?](#) In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents (IVA '24)*, pages 22:1–22:4. Association for Computing Machinery. ArXiv:2409.10357.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. [GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium](#). In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, volume 30. Curran Associates, Inc. ArXiv:1706.08500.
- Marco Jiralerspong, Joey Bose, Ian Gemp, Chongli Qin, Yoram Bachrach, and Gauthier Gidel. 2023. [Feature Likelihood Divergence: Evaluating the Generalization of Generative Models Using Samples](#). In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, volume 36, pages 33095–33119. Curran Associates, Inc. ArXiv:2302.04440.
- Taras Kucherenko, Rajmund Nagy, Youngwoo Yoon, Jieyeon Woo, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2023. [The GENE Challenge 2023: A Large-Scale Evaluation of Gesture Generation Models in Monadic and Dyadic Settings](#). In *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI '23)*, pages 792–801. Association for Computing Machinery. ArXiv:2308.12646.
- Taras Kucherenko, Pieter Wolfert, Youngwoo Yoon, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2024. [Evaluating Gesture Generation in a Large-Scale Open Challenge: The GENE Challenge 2022](#). *ACM Transactions on Graphics*, 43(3):32:1–32:28. ArXiv:2303.08737.
- Rajmund Nagy, Hendric Voss, Youngwoo Yoon, Taras Kucherenko, Teodor Nikolov, Thanh Hoang-Minh, Rachel McDonnell, Stefan Kopp, Michael Neff, and Gustav Eje Henter. 2024. [Towards a GENE Leaderboard – an Extended, Living Benchmark for Evaluating and Advancing Conversational Motion Synthesis](#). *arXiv preprint arXiv:2410.06327*.
- Ali Naseh, Anshuman Suri, Yuefeng Peng, Harsh Chaudhari, Alina Oprea, and Amir Houmansadr. 2025. [Text-to-Image Models Leave Identifiable Signatures: Implications for Leaderboard Security](#). *arXiv preprint arXiv:2510.06525*.
- Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. 2023. [A Comprehensive Review of Data-Driven Co-Speech Gesture Generation](#). *Computer Graphics Forum*, 42(2):569–596. ArXiv:2301.05339.
- Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2018. [Towards Accurate Generative Models of Video: A New Metric and Challenges](#). *arXiv preprint arXiv:1812.01717*.
- Axel Wiebe Werner, Jonas Beskow, and Anna Deichler. 2025. [Gesture Evaluation in Virtual Reality](#). *arXiv preprint arXiv:2509.12816*.
- Yuli Wu, Fucheng Liu, Rüveyda Yilmaz, Henning Konermann, Peter Walter, and Johannes Stegmaier. 2025. [Pragmatic Note on FID for Retinal Image Synthesis](#). *arXiv preprint arXiv:2502.17160*.
- Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. [Speech Gesture Generation from the Tri-modal Context of Text, Audio, and Speaker Identity](#). *ACM Transactions on Graphics*, 39(6):222:1–222:16. ArXiv:2009.02119.
- Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2022. [The GENE Challenge 2022: A Large Evaluation of Data-Driven Co-Speech Gesture Generation](#). In *Proceedings of the 2022 International Conference on Multimodal Interaction (ICMI '22)*, pages 736–747. Association for Computing Machinery. ArXiv:2208.10441.