From Embeddings to Language Models: A Comparative Analysis of Feature Extractors for Text-Only and Multimodal Gesture Generation

Johsac I. G. Sanchez
j216401@dac.unicamp.br
Department of Computer Engineering and Automation,
FEEC - Unicamp
Campinas, SP, Brazil

Paula D. P. Costa
paulad@unicamp.br
Department of Computer Engineering and Automation,
FEEC - Unicamp
Campinas, SP, Brazil

Abstract

Generating expressive and contextually appropriate co-speech gestures is crucial for naturalness in human-agent interaction. While Large Language Models (LLMs) have shown great potential for this task, questions remain regarding the optimal integration of multimodal features and the capabilities of smaller, more accessible models. This study presents a systematic and comparative evaluation of seven gesture generation pipelines, using a robust diffusion-based architecture as our foundation. We investigate the impact of audio (WavLM, Whisper) and text (Word2Vec, Llama-3.2-3B-Instruct) feature extractors to assess the relative contribution of each modality to overall performance. We demonstrate that it is possible to achieve state-of-the-art performance using a significantly smaller LLM (3B parameters) than previous benchmarks, without sacrificing quality. Our results, based on objective metrics and a comprehensive perceptual evaluation, reveal that pipelines incorporating Llama-3.2-3B-Instruct not only outperform references in semantic appropriateness and human-likeness but are also perceived as more appropriate by human evaluators. This work offers guidance for feature and model selection in gesture synthesis, balancing generative quality with model accessibility.

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Human-centered computing** → *Human-computer interaction (HCI).*

Keywords

Gesture generation, robustness, LLM, Multimodal, Diffusion Transformer

ACM Reference Format:
Johsac I. G. Sanchez and Paula D. P. Costa. 2025. From Embeddings to Language Models: A Comparative Analysis of Feature Extractors for Text-Only and Multimodal Gesture Generation. In *Proceedings of the International Workshop on Generation and Evaluation of Non-verbal Behaviour for Embodied Agents (GENEA '25), October 27–28, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3746268.3759431

GENEA '25, Dublin, Ireland

ACM ISBN 979-8-4007-2050-5/2025/10
https://doi.org/10.1145/3746268.3759431

1 Introduction

Human communication is a multimodal symphony where the spoken word is intrinsically intertwined with non-verbal language. Co-speech gestures, in particular, play a crucial role, enriching discourse with emotional nuances, emphasis, and contextual cues that facilitate more natural and understandable interaction [10, 24]. In the realm of conversational agents and virtual character animation, the ability to generate gestures that are not only temporally synchronized with speech but also contextually appropriate and expressive is fundamental to enhancing the realism and quality of human-machine interaction. Historically, automatic gesture generation has predominantly focused on the use of audio features, exploiting prosody and acoustic content to drive the animation [11]. While these approaches have made significant strides in synchronization and rhythm, they often lack the semantic richness that textual content can provide [29].

Recently, there has been a growing interest in incorporating textual information to guide gesture synthesis, recognizing that the meaning and intent behind words are key drivers of gestural behavior [2]. Pioneering works such as LLAniMAtion [29] have demonstrated the potential of Large Language Models (LLMs), like Llama 2 [28], to generate gestures from text features. Notably, these text-driven models are guided by word-level timings derived from the audio, achieving remarkable semantic accuracy and even outperforming audio-only models. These advancements suggest that LLMs can encode deep linguistic structures relevant to gesture generation. However, the focus on large-scale models, such as the 7B-parameter Llama 2 used in LLAniMAtion, introduces challenges related to computational cost, inference speed, and accessibility.

In this context, our work investigates the potential of small-scale LLMs to contribute to the realism of multimodal gesture generation. We examine how these models interact with audio features and different fusion strategies to capture the nuances of timing, expressiveness, and semantic intent in gestures—elements essential to believable human-machine interaction. This exploration is particularly relevant given the practical demand for deployable models in real-time settings, where computational resources are limited. Understanding how far we can push gesture realism under these constraints offers critical insights for building expressive and accessible systems.

To this end, we conduct an exploratory and systematic evaluation of various audio and text feature integration strategies for body gesture synthesis, using the DiffuseStyleGesture+ [30] architecture, a top-performing model from the GENEA Challenge 2023 [14], as our foundation. Rather than introducing a novel architecture, our main contribution lies in the comparative analysis — both objective and perceptual — of multiple generation pipelines. We investigate how different feature extractors influence the quality, the realism, and the appropriateness of the generated gestures. For audio, we explore established extractors like WavLM [4] and, following the same philosophy of using large pre-trained models, the Whisper model [23]. For text, we compare word embedding models like Word2Vec [3] with a more compact 3B-parameter variant of Llama 3.2 [6]. This research seeks to answer the key questions: Are text-only models consistently superior? Can smaller LLMs drive realistic behavior? How text-only approaches compare to fully multimodal approaches?

To address these questions systematically, we build upon the robust DiffuseStyleGesture+ [30] architecture and the GENEA 2023 dataset [14, 16]. Our primary contribution is a systematic and comparative analysis of seven different pipelines, pitting traditional embeddings against modern language models in both text-only and multimodal configurations. Through this evaluation, we make our second key contribution: we demonstrate that a smaller language model (Llama-3.2-3B-Instruct) is not only capable of driving gesture synthesis, but can achieve state-of-the-art results, outperforming established multimodal baselines.

This conclusion is supported by a comprehensive suite of objective metrics—including Fréchet Gesture Distance (FGD), GAC Dice Score, Beat Alignment Score (BAS), and average jerk—as well as a comprehensive perceptual evaluation. Our results extend the findings of previous studies like LLAniMAtion [29] by showing that high-quality, semantically appropriate gestures are not the exclusive domain of massive-scale models; we also offer clear guidance on feature selection for creating more realistic and appropriate virtual agents. The full implementation, including the evaluation framework and all pipeline configurations, is available at: https://github.com/AI-Unicamp/LLM-Gesture-Pipelines.

2 Related Works

The automatic generation of co-speech gestures from speech has been an active area of research for decades, evolving from rule-based systems to increasingly sophisticated data-driven approaches. The recent literature can be organized around the key question this study seeks to answer. The comparison between multimodal using both audio and text features versus those using text-based approaches.

2.1 The Dominant Paradigm: Multimodal Approaches

Recently, the dominant paradigm has been multimodality, based on the premise that combining audio and text features should yield the most robust results [20, 32]. Audio provides prosodic information for the rhythm and synchronization of gestures [11], while text contributes to semantic content [27]. Models like Gesticulator [12] and most systems presented in the GENEA challenges [13–15] have operated under this assumption, fusing audio representations (such as MFCCs or WavLM) with text representations (such as Word2Vec or BERT) to condition their generative models. DiffuseStyleGesture+ itself, which serves as the foundation for our architecture, is an example of this approach, combining multiple audio and text features [5, 30]. These systems represent the state-of-the-art and provide a benchmark that serves as a valuable reference for evaluating new approaches, particularly those that deviate from this audio-text fusion.

2.2 The Rise of Text-Driven Models and the Role of LLM Scale

Recently, a line of research has emerged that challenges the multimodal paradigm. The pioneering work LLAniMAtion [29] showed that features extracted from an LLM (Llama 2) alone were not only sufficient for high-quality gesture generation but also significantly outperformed audio features. Furthermore, they found that combining both modalities offered no substantial improvement over using text exclusively [29]. This finding suggested that powerful LLMs could implicitly capture prosodic and rhythmic information from the context and structure of the language, rendering the audio input redundant. Supporting this idea from a different angle, other works have focused on using LLMs for gesture selection. Research by Hensel et al. [7] and Torshizi et al. [27] has employed GPT-3.5-turbo and GPT-4 to analyze the text of an utterance and suggest appropriate gestures from a repertoire, demonstrating the profound capacity of LLMs for contextual and semantic gestural reasoning.

Although this evidence points towards a potential superiority of text-centric models, the works that established this potential have relied on large-scale models. LLAniMAtion used a 7-billion-parameter version of Llama 2 [29], while gesture selection studies employed the APIs of OpenAI's most powerful models, such as GPT-4 [7, 27]. The use of these massive models, while effective, presents significant barriers in terms of computational cost, accessibility for research, and real-time applicability.

In the gesture generation literature, the performance of smaller LLMs (in the 1-3B parameter range) in an end-to-end generation task is a largely unexplored area. It is unclear whether the semantic richness required for this task is an emergent property that only appears in large-scale models, or if more compact models can be sufficient. The exploration of this question is one of the central contributions of our work.

2.3 The Challenge of Standardized Evaluation

Finally, a robust and standardized evaluation framework is indispensable. As Nagy et al. note in their proposal for a GENEA leaderboard [19], evaluation in gesture generation has historically been fragmented, making it nearly impossible to directly compare results from different publications [19]. Objective metrics often show a low correlation with human perception [15], and subjective studies vary enormously in their design [19]. The GENEA Challenges [13–15] were created to mitigate this problem, providing a controlled ecosystem (same data, visualization, and evaluation protocol) for fair comparison. This paper fully aligns with that philosophy. By using the GENEA 2023 dataset [14, 16] and a rigorous objective and perceptual evaluation protocol. However, for our perceptual study on appropriateness, we adopt the direct pairwise comparison methodology from [29], as it directly measures user preference between competing models. This ensures our conclusions are empirically sound and comparable to related state-of-the-art work.

3 Proposed Method

Our work focuses on the comparative evaluation of seven gesture generation pipelines and one reference. These pipelines were designed to isolate and analyze the impact of three fundamental axes: 1) the audio feature extractor, 2) the text feature extractor, and 3) the fusion and generation architecture.

The pipelines are grouped into four categories to facilitate analysis, as detailed in Table 1. This table serves as a central reference for the nomenclature used throughout the paper.

The general workflow of our experiments is illustrated in Figure 1. All proposals share the same input data sources (Audio and Text) and use the same base diffusion architecture, but they differ in the intermediate "Fusion/Encoding" modules and, in some cases, in the final "Generator" architecture. Each evaluated pipeline represents a unique path through this diagram, allowing us to perform direct and controlled comparisons.

3.1 Dataset

For all our experiments, we utilized the dataset from the GENEA Challenge 2023. This dataset, derived from "Talking With Hands" [14, 16], provides dyadic conversations with 30 fps motion captures in BVH format, synchronized audio, and text transcriptions for both interlocutors. The data is split into a training set containing 17 speakers and a test set with 3 speakers. Although the original dataset contains dyadic interactions, our study focuses on a monadic gesture generation task. Therefore, only the main-agent data were used for training and testing, while the interlocutor data were excluded from the experimental setup.

3.2 Base Generation Architecture

Our work is built upon the DiffuseStyleGesture+ architecture [30], an influential model based on the Motion Diffusion Model (MDM) [25]. To simplify referencing, we will denote the original, unmodified DiffuseStyleGesture+ pipeline as (`Ref-Base`) in this work.

Architectural Terminology: All pipelines in this study employ a diffusion model with a Transformer backbone. To distinguish between the two different conditioning strategies we evaluate, we adopt the following convention:

- **Transformer-based diffusion** to refer to the architecture inherited from DiffuseStyleGesture+, where conditioning information (audio, text, speaker ID) is processed and then *concatenated* with the noisy gesture representation before being fed into the attention blocks.
- **DiT (Diffusion Transformer)** to specifically refer to pipelines that implement the conditioning architecture proposed by Peebles et al. [22]. This approach is characterized by injecting the timestep and context conditions separately into the Transformer blocks via mechanisms like *Adaptive Layer Norm (AdaLN)* and *PerceiverCrossAttention* [8], respectively.

3.3 Training Strategy and Parameters

The original work that serves as the basis for our architecture, DiffuseStyleGesture+, was trained for 1.2 million steps, and its authors provide a pre-trained model at this checkpoint in their GitHub repository [30]. One of the secondary hypotheses of our study is that it is possible to achieve comparable synthetic gesture quality with fewer training steps. To validate this hypothesis, we evaluated different checkpoints of the training process in the `Ref-Base` architecture.

The full 1.2M step training required approximately 14 days on an NVIDIA RTX Quadro 5000 GPU. The performance of the different checkpoints is shown in Figure 2. The model's quality on the test set, measured by both Fréchet Gesture Distance (FGD) and GAC Dice Score associated to the Gesture Area Coverage (GAC) analysis, does not improve monotonically. While the GAC Dice score peaks at 0.78 and stabilizes, the FGD score reaches its best value (15.96) at 900k steps after a significant dip at 540k steps (16.03). However, the marginal improvement in FGD between 540k and 900k steps comes at the cost of nearly doubling the training time. Consequently, we identified the 540k checkpoint as the optimal balance between high performance and computational cost. Based on this finding, all models in our study were trained for a maximum of 540k steps, reducing the average training time to 6 days per model.

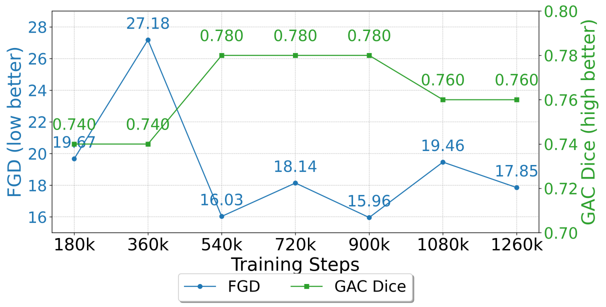

Figure 2: Performance metrics for the Reference Pipeline (`Ref-Base`) at different training checkpoints. The results demonstrate a non-monotonic improvement. The 540k step checkpoint was selected as it offers a strong FGD score (16.03) and the peak GAC Dice score (0.78), representing the best trade-off between model quality and training resources.

3.4 LLM Text Feature Extraction

For pipelines incorporating advanced semantic information (those using Llama-3.2-3B-Instruct), we developed a sophisticated pipeline to extract and temporally align embeddings from a Large Language Model (LLM). Unlike static embeddings like Word2Vec, this approach generates rich contextual representations for each word. The process is as follows:

(1) **Transcription Segmentation:** Transcriptions, provided as lists of words with timestamps (TSV format), are first grouped into coherent sentences or speech units. Segmentation is performed by detecting significant pauses (greater than 1.0 second) between words or by the presence of final punctuation marks.
(2) **Tokenization and Offset Mapping:** Each reconstructed sentence is processed with the LLM's tokenizer, which generates not only tokens but also an `offset_mapping`. This mapping is crucial as it links each token to its exact character position (start and end) within the sentence string.

Table 1: Summary of gesture generation pipelines, divided into reference and proposed pipelines. Proposed pipelines are grouped by category. The table details input components (audio and text) and encoding/generation architecture.

Category	Pipeline ID	Text Embedding (Dim)	Audio Embedding (Dim)	Encoding / Architecture
Reference Pipeline	`Ref-Base`	Word2Vec (300) + flags (2)	WavLM (1024) + acoustic feat (109)	Concatenation + Linear Projection/ Transformer-based diffusion
Proposed Pipelines	`Basic-Whisper`	Word2Vec (300) + flags (2)	Whisper (1280)+ onset (1)	Concatenation + Linear Projection/ Transformer-based diffusion
	`Multi-Fusion`	Llama 3.2 (3072) + flags (2)	WavLM (1024) + acoustic feat (109)	Concatenation + MLP / Transformer-based diffusion
	`Multi-Dual`	Llama 3.2 (3072) + flags (2)	WavLM (1024) + acoustic feat (109)	Dual encoders + Latent fusion / Transformer-based diffusion
	`Multi-Whisper`	Llama 3.2 (3072) + flags (2)	Whisper (1280)+ onset (1)	Dual encoders + Latent fusion / Transformer-based diffusion
	`Text-Only`	Llama 3.2 (3072) + flags (2)	—	Text encoder/ Transformer-based diffusion
	`Multi-DiT`	Llama 3.2 (3072) + flags (2)	WavLM (1024) + acoustic feat (109)	DiT with Perceiver Cross-Attention (from `Multi-Dual`)
	`Text-DiT`	Llama 3.2 (3072) + flags (2)	—	DiT with Perceiver Cross-Attention (from Text-Only)

flags (2) = laughter (1) + onset (1) , acoustic feat (109)= MFCC (40) + Spectrum (64) + Prosody (4) + Onset (1)

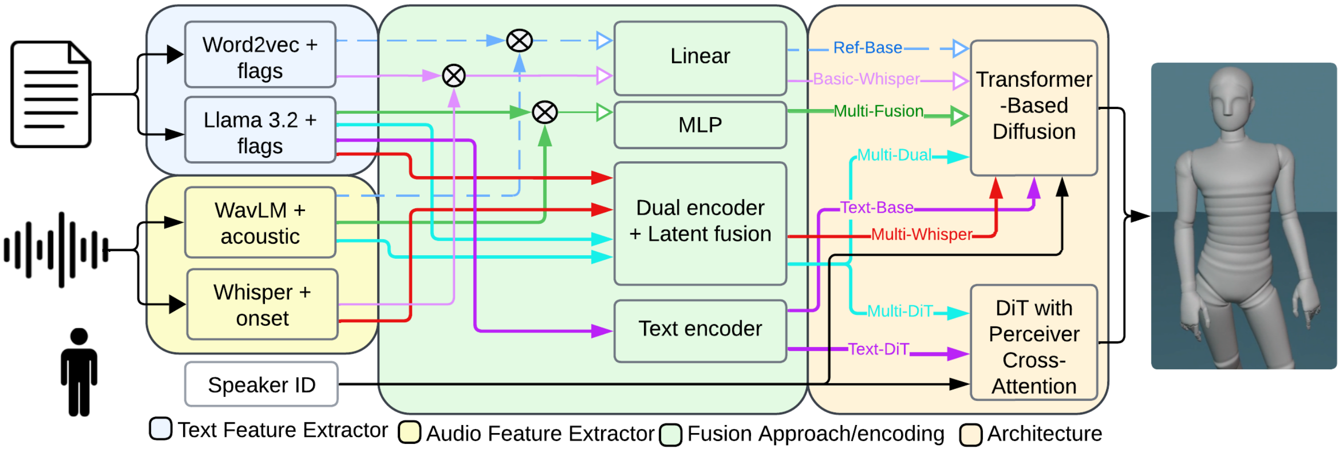

Figure 1: Flowchart of the evaluated experimental pipelines. Each pipeline represents a unique path from the data sources (left), through a Fusion/Encoding module (center), to a Generator architecture (right). This structure allows for a systematic comparison of the impact of each component on gesture generation.

(3) **Contextual Embedding Generation:** The hidden states from the LLM's last layer are obtained for the entire sentence.
(4) **Word-Token Alignment:** For each word from the original transcription, its character range is used to identify all LLM tokens whose offsets overlap with it. This step effectively handles cases where a word is split into multiple sub-tokens.
(5) **Embedding Pooling:** The final embedding for a word is calculated by averaging the hidden states of all tokens that overlap with it. This pooling ensures that the word's representation is robust and informed by its full context in the sentence.
(6) **Feature Track Creation:** Finally, the contextual embedding of each word is assigned to the corresponding 30fps frames of the sequence, using the original timestamps from the TSV file. This results in a dense textual feature track temporally aligned with audio and gesture.

This method ensures that the diffusion model receives information not only about *what* was said but also about *how* and *when*, capturing the flow and structure of spoken language.

3.5 Evaluated Pipelines

To conduct our comparative study, we designed and evaluated eight distinct pipelines, which are detailed in Table 1 and visualized in Figure 1. These pipelines are grouped into four main categories, and their specific conditioning mechanisms are detailed below.

3.5.1 Reference Pipelines. This group establishes the baselines for our study using a simple fusion method.

`Ref-Base`: Our primary reference, a faithful implementation of the DiffuseStyleGesture+ model [30]. Its fusion strategy consists of concatenating the audio (1133 dims) and Word2Vec text (302 dims) features, followed by a single linear projection to reduce the combined vector to a latent representation (1435 ⟶ 128).

`Basic-Whisper`: This alternative baseline maintains the same structure but replaces the audio features with embeddings from the Whisper encoder. The concatenated vector is projected to the latent space (1583 ⟶ 128). This allows for a direct comparison of audio extractors.

3.5.2 Multimodal LLM Pipelines. This category explores different strategies for combining the rich semantic features from Llama 3.2 with audio information.

`Multi-Fusion`: This pipeline tests a "deep fusion" approach. The Llama (3074 dims) and WavLM-based audio (1133 dims) features are first concatenated. Then, a deep MLP performs a non-linear dimensionality reduction on the combined vector (4207 ⟶ 1024 ⟶ 512 ⟶ 128).

`Multi-Dual`: This pipeline employs a "latent fusion" strategy with specialized encoders. It uses two separate MLPs to process each modality independently—one for audio (1133 ⟶ 512 ⟶ 128) and one for text (3074 ⟶ 1500 ⟶ 512 ⟶ 128). Their resulting 128-dimensional latent representations are then concatenated to form the final 256-dimensional conditioning vector.

`Multi-Whisper`: This pipeline replicates the dual-encoder strategy from `Multi-Dual` but substitutes WavLM with Whisper audio features (1281 dims), which are processed by their own audio encoder path (1281 ⟶ 512 ⟶ 128).

3.5.3 Text-Only LLM Pipeline. This pipeline is designed to test the "text is all you need" hypothesis.

`Text-Only`: This pipeline generates gestures using only Llama 3.2 features (3074 dims). The features are processed by a single deep MLP encoder (3074 ⟶ 1500 ⟶ 512 ⟶ 256) that feeds the base Transformer-based Difussion architecture.

3.5.4 Diffusion Transformer (DiT) Architecture Pipelines. This final group evaluates the impact of replacing the base generator with a more advanced Diffusion Transformer (DiT) architecture [22].

`Multi-DiT`: This pipeline applies the DiT architecture to the 256-dimensional multimodal context generated by the `Multi-Dual` pipeline's dual encoders.

`Text-DiT`: This pipeline combines the DiT generator with the 256-dimensional text-only context from the `Text-Only` pipeline encoder.

4 Evaluation

To assess the performance of our gesture generation pipelines, we conducted a comprehensive evaluation divided into two parts. First, we employed a set of objective metrics to quantitatively measure different aspects of the generated motion, such as distributional similarity, spatial coverage, smoothest motion and rhythmic alignment. Second, we performed a perceptual study to understand how these objective measures translate to human experience, focusing on perceived human-likeness and contextual appropriateness.

4.1 Objective Metrics

4.1.1 Fréchet Gesture Distance (FGD) [31]. FGD is a standard metric for evaluating the similarity between the distribution of generated and real gestures. It is calculated by extracting features from a pre-trained autoencoder for both sets of motions (generated and real). The distance is then computed as:

$$\mathrm{FGD} = ||\mu_r - \mu_g||^2 + \mathrm{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}) \tag{1}$$

where μ_r and μ_g are the feature means, and Σ_r and Σ_g are their covariance matrices for the real (r) and generated (g) gestures, respectively. A lower FGD indicates greater similarity between the distributions.

4.1.2 GAC Dice Score. To measure the similarity in the use of gestural space, we use the Dice Score on the Gesture Area Coverage (GAC), as proposed in [26]. The GAC is defined as the union of all poses in a sequence, rasterized into a 2D image to create a coverage map (grid). Given the GAC of real gestures (GAC_r) and generated gestures (GAC_g), the Dice Score is defined as:

$$\mathrm{Dice} = \frac{2 \times |GAC_r \cap GAC_g|}{|GAC_r| + |GAC_g|} \tag{2}$$

It quantifies the overlap between two sets, penalizing both the area that the generated gesture did not cover (false negatives) and the area that it overcovered (false positives). A value close to 1 indicates high spatial overlap, i.e., more expressive and similar generated gestures to real gestures, while a value close to 0 indicates the opposite.

4.1.3 Beat Alignment Score (BAS) [17]. BAS measures the rhythmic synchrony between speech and gesture. It compares the timing of velocity peaks in hand movements with the timing of speech beats, which are extracted from the audio signal. A higher BAS indicates that the gestures are better aligned with the rhythm of the speech. It is formally defined as:

$$\mathrm{BAS} = \frac{1}{N}\sum_{i=1}^{N}\exp\left(-\frac{\left(\min_j |t_i^a - t_j^g|\right)^2}{2\sigma^2}\right) \tag{3}$$

where $\{t_1^a, ..., t_N^a\}$ is the set of N audio beat timestamps, and $\{t_j^g\}$ is the set of gesture beat timestamps (velocity peaks). For each audio beat t_i^a, the score is calculated based on the time difference to the nearest gesture beat, penalized by a Gaussian kernel with a tolerance parameter σ.

4.1.4 Jerk (JM) [18]. Jerk is the third temporal derivative of the joint positions and is used to quantify the smoothness of the motion. A lower Jerk indicates a smoother movement. It is calculated as the sum of the magnitudes of the changes in acceleration throughout the sequence:

$$\mathrm{JM} = \sum_{t} \left| \frac{d^3 P(t)}{dt^3} \right| \tag{4}$$

where $P(t)$ is the position of the joints at time t.

4.2 Objective Results

The objective evaluation of our pipelines focused on four key metrics. The results, obtained after training all models for 540k steps, are summarized in Table 2.

Table 2: Objective metric results for the proposed pipelines, trained for 540k steps.

Pipeline ID	FGD (↓)	BAS (↑)	Jerk (↓)	GAC Dice (↑)
Ref-Base	16.03	0.74	**0.24**	**0.78**
Basic-Whisper	9.13	0.02	0.66	0.75
Multi-Fusion	11.20	**0.76**	0.49	**0.78**
Multi-Dual	10.81	0.75	0.47	**0.78**
Multi-Whisper	9.43	0.02	0.87	0.76
Text-Only	**8.98**	**0.76**	0.60	**0.78**
Multi-DiT	16.24	**0.76**	0.70	0.76
Text-DiT	9.05	**0.76**	0.54	0.75

The **Fréchet Gesture Distance (FGD)**, which measures distributional similarity to real gestures, reveals the most significant differences. The results clearly position the Llama 3.2-based strategies as the most effective. Notably, the `Text-Only` pipeline, which relies solely on textual input, achieved the best performance with an FGD of **8.98**. This is closely followed by the `Text-DiT` (9.05) and the Whisper-based pipelines (9.13 and 9.43). This represents a remarkable improvement of over 44% compared to the `Ref-Base` multimodal pipeline (16.03) at the same training checkpoint. This finding strongly suggests that high-quality textual representations can be more effective than standard multimodal fusion for achieving high-fidelity gesture generation, and can do so efficiently.

The **GAC Dice**, evaluating spatial overlap with ground-truth gestures. The top-performing pipelines, including `Text-Only`, `Multi-Fusion`, and `Multi-Dual`, all achieved a high score of **0.78**, indicating that their generated gestures occupy a gestural space very similar to that of human speakers.

The **Beat Alignment Score (BAS)** presented the most polarized results. All pipelines using WavLM or Llama 3.2 without Whisper maintained a consistently good BAS in the 0.74-0.76 range. The fact that the textual `Text-Only` pipeline, which receives word timings from the audio transcript but no other acoustic features, achieves a high BAS of 0.76 is a key finding. Conversely, both pipelines using Whisper features (`Basic-Whisper` and `Multi-Whisper`) yielded an extremely low BAS of **0.02**, indicating a lack of alignment with speech prosodic beats, which may primarily affect rhythmic gestures, though not necessarily semantic ones.

Finally, the **Jerk (JM)** metric revealed an interesting trend. The reference pipeline produced the smoothest motion (JM of 0.24). In contrast, all pipelines incorporating the Llama 3.2 LLM produced significantly higher jerk values (ranging from 0.47 to 0.87), indicating less smooth, more dynamic movements. The implications of this are explored in the Discussion section 5.

Based on these objective results, we selected the following representative subset of four conditions for the perceptual study to test our core hypotheses:

1. **`Text-Only`**: Chosen as the best-performing pipeline in objective metrics (especially FGD), to test the "text-only" hypothesis.
2. **`Text-DiT`**: Selected to directly compare the impact of the DiT architecture against its non-DiT counterpart, isolating the effect of the generator architecture on the same text-only condition.
3. **`Ref-high`** (`Ref-Base` @540k steps): Included as a high-quality baseline, representing the best performance of the reference pipeline.
4. **`Ref-low`** (`Ref-Base` @180k steps): Included as a low-quality anchor, a standard practice in MUSHRA-type tests.

4.3 Perceptual Evaluation

The perceptual evaluation was designed to capture human perception of gesture quality and was conducted with 44 volunteer participants recruited among University students. For this study, we generated 12-second video clips for each evaluated stimulus. The average time to complete the evaluation was 29 minutes. The study was divided into two parts:

4.3.1 Part 1: Naturalness (Human-likeness). The objective of this part was to measure the perceived human-likeness. The HEMVIP evaluation [9] was used , based on the MUSHRA protocol [1]. In each of the 8 trials, participants rated four videos (one for each pipeline) corresponding to a unique speech segment, on a continuous 0-100 scale, with semantic anchors such as "bad" (0-20), "poor" (21-40), "fair" (41-60), "good" (61-80) and "excellent" (81-100).

4.3.2 Part 2: Appropriateness. This part focused on evaluating contextual appropriateness and synchrony using a pairwise comparison (A/B test), following the methodology of LLAniMAtion [29]. This method directly compares two independent models (e.g., model X vs. model Y), differing from the GENEA challenge's mismatching paradigm [15]. In each of the 24 unique trials, participants viewed two videos and indicated their preference for which was better in terms of synchrony, disregarding pure naturalness, on a 5-point scale (-2 to +2) [21]: left is clearly better (-2), Left slightly better (-1), Both equal (0), Right slightly better (+1) and right is clearly better (+2). We then calculated the Mean Appropriateness Score (MAS) from the absolute value of these scores, providing a measure of the preference strength for each pipeline.

Both parts of the study employed full randomization of the stimulus order and video positions to minimize bias.

4.4 Perceptual Evaluation Results

The results for both human-likeness and appropriateness are summarized in Figures 3, 4, and detailed in Table 3.

Table 3: Summary of perceptual evaluation results. Values represent Mean ± 95% Confidence Interval. Superscript letters (^{a,b}) denote statistical significance groups based on a post-hoc Tukey HSD test ($\alpha = 0.05$). Models not sharing a common letter are significantly different from each other within the same metric.

Pipeline	Human-likeness	Appropriateness (MAS)
`Text-Only`	57.21 ± 2.62^{a}	1.34 ± 0.07^{a}
`Text-DiT`	64.13 ± 2.37^{b}	1.24 ± 0.08^{a}
`Ref-high`	59.65 ± 2.42^{ab}	0.62 ± 0.09^{b}
`Ref-low`	57.20 ± 2.44^{a}	0.61 ± 0.10^{b}

For **Human-likeness**, the `Text-DiT` pipeline was perceived as the most natural, achieving the highest mean score (64.13). A one-way ANOVA confirmed a significant difference between the models ($p < 0.001$). Post-hoc analysis revealed that `Text-DiT` was rated as significantly more human-like than both `Text-Only` (p=0.0006) and the low-quality reference `Ref-low` (p=0.0006). This suggests that the DiT architecture plays a crucial role in improving the perceived kinematic quality of the generated motion.

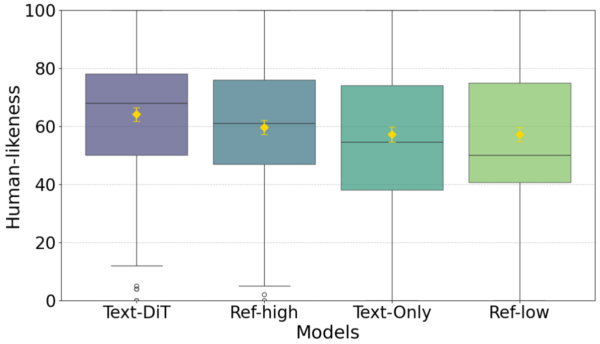

Figure 3: Boxplot of human-likeness scores from the MUSHRA evaluation for each of the selected pipelines. The line within the box is the median, and circles are outliers. The yellow diamond represents the mean score for each condition. Higher scores indicate greater perceived naturalness.

For **Appropriateness**, the results show a preference for LLM-based pipelines. Both `Text-Only` (MAS=1.34) and `Text-DiT` (1.24) were rated as significantly more appropriate than both `Ref-high` (0.62) and `Ref-low` (0.61) ($p < 0.001$ for all comparisons). Table 4 details the win rates in these pairwise comparisons, offering a more granular view. This breakdown reveals the extent of the LLM models' superiority: `Text-Only` was preferred over the high-quality reference (`Ref-high`) in 72.2% of trials, with only 18.8% of evaluators choosing the reference. A similar dominance is observed for `Text-DiT` against the same reference (67.6% wins). In contrast, the comparison between the two LLM-based pipelines is much more balanced (`Text-Only` winning 45.5% vs. `Text-DiT` winning 39.8%), suggesting that while both are perceived as highly appropriate, neither has a definitive edge in these comparisons. These results provide strong perceptual validation of the power of LLMs to generate gestures that are not just kinematically plausible, but semantically and contextually appropriate.

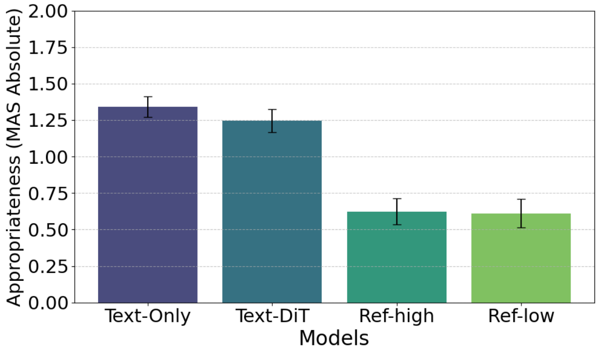

Figure 4: Mean Appropriateness Scores (MAS) absolute with 95% Confidence Intervals from the AB pairwise comparison tests for each of the selected pipelines. Higher scores indicate greater perceived appropriateness.

Table 4: Pairwise Comparison Results for Appropriateness: Percentage of Wins (Left vs. Right) and Ties.

Pipeline	Win Left	Tie	Win Right	Pipeline
`Text-Only`	45.5%	14.8%	39.8%	`Text-DiT`
`Text-Only`	72.2%	9.1%	18.8%	`Ref-high`
`Text-Only`	73.9%	8.0%	18.2%	`Ref-low`
`Text-DiT`	67.6%	13.6%	18.8%	`Ref-high`
`Text-DiT`	66.5%	18.2%	15.3%	`Ref-low`
`Ref-high`	29.5%	44.3%	26.1%	`Ref-low`

5 Discussion

Our results, which combine objective metrics and rigorous perceptual evaluation, offer a nuanced view of gesture generation. A central finding is the resounding success of text-driven strategies, particularly those using Llama 3.2-3B. The `Text-Only` and `Text-DiT` pipelines not only dominated key objective metrics such as FGD and GAC Dice, but also were perceived by participants as significantly more appropriate than reference pipelines. This strongly corroborates the thesis of LLAniMAtion [29] that rich semantic encodings of text are a more potent driver for the contextual appropriateness of gestures than audio features alone. Our work extends this finding by demonstrating that this level of performance can be achieved with a considerably smaller LLM (3B), challenging the notion that top-tier gesture generation is exclusive to massive-scale models.

A particularly interesting observation arises from comparing the Jerk (JM) metric with the perception of appropriateness. Table 5 compares these metrics for the perceptually evaluated pipelines.

Our LLM-based pipelines exhibited higher Jerk values than the references, objectively indicating less smooth movement. However,

Table 5: Comparison of Jerk (Objective) and Mean Appropriateness Score (MAS, Subjective).

Pipeline	Jerk (↓)	MAS (↑)
`Text-Only`	0.60	1.34
`Text-DiT`	0.54	1.24
`Ref-high`	0.24	0.62
`Ref-low`	0.20	0.61

these same pipelines were rated as significantly more appropriate. To validate the choice of a parametric correlation test, we first assessed the data for normality using the Shapiro-Wilk test. The results indicated that Jerk (W = 0.84, p = 0.21) and MAS (W = 0.79, p = 0.09) significantly met the normality assumption. A Pearson correlation test was conducted. The analysis revealed a strong, statistically significant positive correlation (r = 0.997, p < 0.01), quantitatively confirming that higher jerk is associated with higher perceived appropriateness in our study. Furthermore, qualitative feedback from the 44 volunteers described gestures from the LLM pipelines as "very expressive", in stark contrast to the references, often described as "unexpressive" or "robotic". This challenges the simplistic interpretation that lower Jerk is always preferable, as minimal jerk can also correspond to static or lifeless motion. In fact, this disconnect between objective smoothness metrics and human perception has been previously observed. The GENEA Challenge 2022 [15], for instance, found that most objective metrics, including average jerk, were not well aligned with subjective human-likeness ratings. In that study, some systems with objectively "good" smooth motion were perceived as less human-like by evaluators. Our work reinforces this finding, suggesting that for expressiveness and appropriateness, higher jerk might even positively correlate with more dynamic, varied, and energetic gestures, which humans perceive as more contextually relevant, rather than merely "noisy" or erratic.

In terms of human-likeness, the `Text-DiT` pipeline was the clear winner. This suggests that while the LLM provides the crucial semantic "what," the generator's architecture determines the kinematic "how." The global attention mechanism of the Diffusion Transformer (DiT) may be inherently better at modeling the complex, long-range dependencies of whole-body kinematics, resulting in motion that is perceived as more coordinated and natural.

Finally, the low Beat Alignment Score (BAS) Whisper-based pipelines warrants careful consideration. A plausible hypothesis is that Whisper, optimized for automatic speech recognition (ASR) [23], excels at extracting phonetic content but may abstract away fine-grained prosodic variations crucial for tight rhythmic synchrony in beat gestures. In contrast, models like WavLM, pretrained on raw waveforms, may better preserve this information. However, given that the `Multi-Whisper` pipeline still achieved a competitive FGD score, it appears the strong semantic guidance from the LLM can partially compensate for rhythmic deficiencies. This underscores that a single metric is insufficient to capture overall quality. The videos generated from each pipeline will be published on our GitHub, allowing for a qualitative assessment of whether this low BAS score is perceptually salient.

6 Limitations

While our study was designed for systematic comparison, it has several limitations that open avenues for future work.

Although our study is systematic, it does not cover all possible design combinations. For instance, we did not test Whisper features with a DiT architecture, nor did we explore more complex or attention-based fusion mechanisms beyond concatenation and latent-space fusion. This leaves room for future work to investigate other promising architectural configurations.

The poor performance of Whisper-based pipelines in rhythmic alignment is a limitation of our specific configurations. It does not rule out the possibility that Whisper could be effective with different integration methods, fine-tuning, or if its features were used to supplement a model with an already strong rhythmic base.

7 Conclusion and Future Work

In this work, we conducted a systematic evaluation of seven gesture generation pipelines, focusing on the trade-offs between input modalities, feature extractors, and architectural choices. Our findings provide several key insights. First, we demonstrate that text-driven pipelines using a smaller LLM (Llama-3.2-3B-Instruct) can significantly outperform traditional audio-text models, particularly in perceived semantic appropriateness. This confirms that the rich contextual understanding of LLMs is a powerful driver for gesture generation and that this capability is not exclusive to massive-scale models. Second, our analysis reveals a statistically significant correlation between objective metrics and human perception, where higher motion jerk can correlate with greater perceived expressiveness, challenging the conventional wisdom that smoothness is always optimal. Finally, we show that advanced generator architectures such as the Diffusion Transformer can further enhance perceived naturalness. This work provides a clear guide for model and feature selection in gesture synthesis.

Future work could move beyond simple concatenation and explore more sophisticated, attention-based fusion architectures. Such models could learn to dynamically weigh audio and text features, potentially combining the rhythmic strengths of audio with the semantic power of text more effectively. Other future work could investigate how to condition generative models not just on content, but also on stylistic parameters, allowing animators or users to dial between "smooth and reserved" and "energetic and expressive" motion. To facilitate reproducibility and encourage further research, the code and trained models, and generated videos for this study are publicly available on our GitHub repository: https://github.com/AI-Unicamp/LLM-Gesture-Pipelines.

Acknowledgments

This study was partially funded by the Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior – Brasil (CAPES) – Finance Code 001. This project was supported by the Ministry of Science, Technology, and Innovations, with resources from Law No. 8.248, of October 23, 1991, under the PPI-SOFTEX program, coordinated by Softex and published as Cognitive Architecture (Phase 3), DOU 01245.003479/2024-10.The authors of this research are also affiliated to the Artificial Intelligence Lab., Recod.ai.

References

[1] 2015. *Method for the subjective assessment of intermediate quality level of audio systems.* Recommendation BS.1534-3. International Telecommunication Union, Radiocommunication Sector (ITU-R), Geneva, Switzerland. Also known as the MUSHRA test protocol.

[2] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. 2021. Text2Gestures: A Transformer-Based Network for Generating Emotive Body Gestures for Virtual Agents. In *2021 IEEE virtual reality and 3D user interfaces (VR)*. IEEE, 1–10.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the association for computational linguistics* 5 (2017), 135–146.

[4] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (2022), 1505–1518.

[5] Johsac Isbac Gomez Sanchez, Kevin Adier Inofuente Colque, Leonardo Boulitreau de Menezes Martins Marques, Paula Dornhofer Paro Costa, Rodolfo Luis Tonoli, et al. 2024. Benchmarking Speech-Driven Gesture Generation Models for Generalization to Unseen Voices and Noisy Environments. In *COMPANION PUBLICATION OF THE 26TH INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION, ICMI 2024 COMPANION.* 5.

[6] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[7] Laura Birka Hensel, Nutchanon Yongsatianchot, Parisa Torshizi, Elena Minucci, and Stacy Marsella. 2023. Large language models in textual analysis for gesture selection. In *Proceedings of the 25th International Conference on Multimodal Interaction.* 378–387.

[8] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In *International conference on machine learning.* PMLR, 4651–4664.

[9] Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, and Gustav Eje Henter. 2021. HEMVIP: Human evaluation of multiple videos in parallel. In *Proceedings of the 2021 International Conference on Multimodal Interaction.* 707–711.

[10] Adam Kendon. 1994. Do Gestures Communicate? A Review. *Research on language and social interaction* 27, 3 (1994), 175–200.

[11] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing Input and Output Representations for Speech-Driven Gesture Generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents.* 97–104.

[12] Taras Kucherenko, Patrik Jonell, Sanne Van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 international conference on multimodal interaction.* 242–250.

[13] Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2021. A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA Challenge 2020. In *Proceedings of the 26th International Conference on Intelligent User Interfaces.* 11–21.

[14] Taras Kucherenko, Rajmund Nagy, Youngwoo Yoon, Jieyeon Woo, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2023. The GENEA Challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In *Proceedings of the 25th international conference on multimodal interaction.* 792–801.

[15] 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 GENEA Challenge 2022. *ACM Transactions on Graphics* 43, 3 (2024), 1–28.

[16] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S Srinivasa, and Yaser Sheikh. 2019. Talking With Hands 16.2M: A Large-Scale Dataset of Synchronized Body-Finger Motion and Audio for Conversational Motion Analysis and Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 763–772.

[17] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021. AI Choreographer: Music Conditioned 3D Dance Generation With AIST++. In *Proceedings of the IEEE/CVF international conference on computer vision.* 13401–13412.

[18] Pietro Morasso. 1981. Spatial control of arm movements. *Experimental brain research* 42, 2 (1981), 223–227.

[19] 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 GENEA Leaderboard–an Extended, Living Benchmark for Evaluating and Advancing Conversational Motion Synthesis. *arXiv preprint arXiv:2410.06327* (2024).

[20] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. 2023. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. In *Computer Graphics Forum*, Vol. 42. Wiley Online Library, 569–596.

[21] Etienne Parizet, Nacer Hamzaoui, and Guillaume Sabatie. 2005. Comparison of some listening test methods: a case study. *Acta Acustica united with Acustica* 91, 2 (2005), 356–364.

[22] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision.* 4195–4205.

[23] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning.* PMLR, 28492–28518.

[24] Michael Studdert-Kennedy. 1994. Hand and Mind: What Gestures Reveal About Thought. *Language and Speech* 37, 2 (1994), 203–209.

[25] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. 2022. Human motion diffusion model. *arXiv preprint arXiv:2209.14916* (2022).

[26] Rodolfo Luis Tonoli, Paula Dornhofer Paro Costa, Leonardo Boulitreau de Menezes Martins Marques, and Lucas Hideki Ueda. 2024. Gesture Area Coverage to Assess Gesture Expressiveness and Human-Likeness. In *Companion Proceedings of the 26th International Conference on Multimodal Interaction.* 165–169.

[27] Parisa Ghanad Torshizi, Laura B Hensel, Ari Shapiro, and Stacy C Marsella. 2025. Large Language Models for Virtual Human Gesture Selection. *arXiv preprint arXiv:2503.14408* (2025).

[28] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[29] Jonathan Windle, Iain Matthews, and Sarah Taylor. 2024. Llanimation: Llama Driven Gesture Animation. In *Computer Graphics Forum*, Vol. 43. Wiley Online Library, e15167.

[30] Sicheng Yang, Haiwei Xue, Zhensong Zhang, Minglei Li, Zhiyong Wu, Xiaofei Wu, Songcen Xu, and Zonghong Dai. 2023. The DiffuseStyleGesture+ entry to the GENEA Challenge 2023. In *Proceedings of the 25th International Conference on Multimodal Interaction.* 779–785.

[31] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–16.

[32] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. 2023. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 4 (2023), 2430–2449.