DOI: 10.1111/cgf.15167

ACM SIGGRAPH / Eurographics Symposium on Computer Animation 2024
M. Skouras and H. Wang
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 43 (2024), Number 8

LLAniMAtion: LLAMA Driven Gesture Animation

J.Windle[1], I. Matthews[1] and S. Taylor[2]

[1]University of East Anglia, United Kingdom
[2]Independent Researcher, United Kingdom

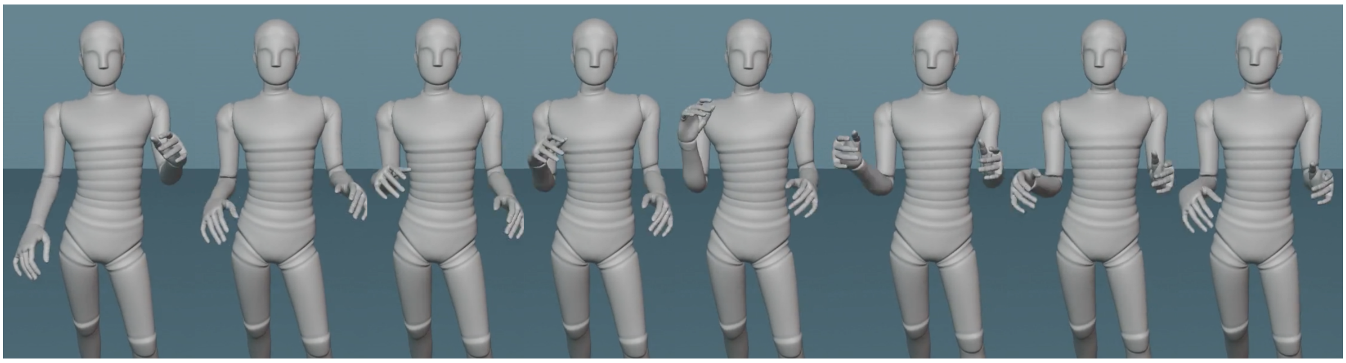


Figure 1: *We present a method for generating gestures from speech using* LLAMA2 *features derived* from text *as the primary input, producing temporally aligned and contextually accurate gesture animation. This sequence generated from our method shows a speaker mimicking the use of a fork with their right hand while describing eating crab.*

Abstract

Co-speech gesturing is an important modality in conversation, providing context and social cues. In character animation, appropriate and synchronised gestures add realism, and can make interactive agents more engaging. Historically, methods for automatically generating gestures were predominantly audio-driven, exploiting the prosodic and speech-related content that is encoded in the audio signal. In this paper we instead experiment with using Large-Language Model (LLM) features for gesture generation that are extracted from text using LLAMA2*. We compare against audio features, and explore combining the two modalities in both objective tests and a user study. Surprisingly, our results show that* LLAMA2 *features on their own perform significantly better than audio features and that including both modalities yields no significant difference to using* LLAMA2 *features in isolation. We demonstrate that the* LLAMA2 *based model can generate both beat and semantic gestures without any audio input, suggesting LLMs can provide rich encodings that are well suited for gesture generation.*

CCS Concepts

• ***Computing methodologies*** → ***Machine learning algorithms;*** *Animation;*

1. Introduction

Co-speech gesturing plays a crucial role in communication, as gestures effectively convey emotions and emphasis and enhance interactions by introducing social and contextual cues. These cues contribute to increased understanding, improved turn-taking, and enhanced listener feedback [Ken94, McN85, SK94, DRBD12]. Gestures have been found to influence what a listener hears [BP21], emphasising the importance of accurately depicting body motion during speech in applications such as video production, video games, avatars, virtual agents, and robotics. The ability to automatically produce realistic gestures from speech has broad applications in these areas.

There is a co-dependency between speech and gesture, where gesture production is a complex function of the speech content, semantics and prosody. For instance, beat gestures synchronise with the timing of the speech audio dynamics, while iconic gestures convey the shape of the discussed topic [BP21]. Current research often focuses on speech-to-gesture generation using audio features as the primary modality. While audio features are effective in encoding prosody, they may not capture semantics as well. Conversely, text features capture the content, but may lack prosodic information. It becomes apparent that a combination of features may yield optimal results.

Large-Language Models (LLMs) are exposed to large natural language corpora, making them exceptional in language and content understanding. In this paper we explore the integration of LLM embeddings into a gesture generation model to improve the semantic accuracy of co-speech gestures. We experiment with methodologies for combining LLM embeddings with audio features, and report the objective and perceptual performance to determine the contribution of each feature. Surprisingly, our results show that LLM features on their own perform significantly better than audio features, and no significant difference is recorded when these two modalities are used in combination. Our approach, called LLAniMAtion, utilises LLAMA2 language embeddings [TMS*23] and optionally combines them with Problem Agnostic Speech Encoding (PASE+) audio features [RZP*20] in a Transformer-XL architecture [WMMT23]. We show that our LLM-based LLAniMAtion produces gestures that exhibit varied motions, capturing both beat and semantic gestures. Our key contributions can be summarised as follows:

- We are one of the first to integrate LLM features into a gesture animation model.
- We evaluate the performance impact of using LLAMA2 features in combination with audio features and in isolation.
- We demonstrate that LLM features contribute more to the perceived quality of the resulting gesture animations than audio features.

2. Related Work

Deep learning approaches to speech-driven gesture generation have historically relied on audio features as their primary input. While text-based features have gained momentum in recent research, the utilisation of LLM features remains limited. We review features used in gesture generation and on LLM models used in gesture and related fields.

2.1. Speech Features for Gesture Generation

Gesture generation systems widely adopt audio-based features. In the review by Nyatsanga et al. [NKA*23] out of 40 methods reviewed, 35 used audio as an input feature. In contrast, only 17 methods involved text as an input. Audio features can be embedded using various methods. Perhaps most common is the use of Mel-frequency Cepstral Coefficient (MFCC) [HKS*18, AHKB20, QTZ*21, HES*22, PKJS20, ANBH23], sometimes combined with other prosodic features such as pitch (F0) and energy [KHH*19]. Other latent representations such as Wav2Vec 2.0 [BZMA20] and PASE+ [RZP*20] have grown in popularity as these can also effectively encode important speech-related information as well as prosodic features [WGT22, WMMT23, NRB*24], while improving speaker independence of the representation. Audio features are advantageous with regard to beat gesture performance as these have a close relationship to prosodic activity, such as acoustic energy and pitch [WTGM22, PHEGD20].

Numerous approaches leverage a combination of both audio and text features, with different methods for incorporating textual information. Word rhythm was used by [ZBC22] where words are encoded in a binary fashion, taking the value 1 if a word is spoken and 0 if not. Other works, such as those by Windle et al. [WGT22, WMMT23] and Yoon et al. [YCL*20] integrate FastText embeddings and [BGJM17] which extend the Word2Vec approach [wor13] exploiting sub-word information. BERT [DCLT18] features have been successfully used in conjunction with audio in the work of Ao et al. [AGL*22]. BERT, originally designed for language modelling and next-sentence prediction, is composed of transformer encoder layers. Kucherenko et al. [KNN*21] also found that it is possible to predict gesture properties related to semantic gesture meaning from FastText embeddings alone, but not from prosodic audio features alone. On the other hand, rhythm-related gesture properties are better predicted from audio features than text.

Using text as the exclusive input for gesture generation is infrequent, and performance is often limited when used. Yoon et al. [YKJ*19] and Bhattacharya et al. [BRB*21] employ GloVe word embedding vectors [PSM14] to facilitate gesture generation.

Despite the recognised advantages of text-based features, to the best of our knowledge, LLMs have not been used in the context of gesture generation, whether in isolation or in combination with audio inputs. This highlights a gap in the current research landscape that we aim to explore in this paper.

2.2. Large Language Models

Given the close relationship between language and gesture, the recent advances in LLM performance present a promising avenue for advancing gesture generation. We provide a brief overview of LLM approaches and refer the reader to Yang et al. [YJT*23] for a comprehensive review.

LLM approaches fall into two categories: Encoder-Decoder/ Encoder only and Decoder only, often referred to as Bidirectional Encoder Representations from Transformers (BERT) [DCLT18] and Generative Pre-trained Transformer (GPT) style, respectively. These models typically exhibit a task-agnostic architecture. Our primary focus in this work is on GPT-style models, which currently stand as leaders in LLM performance. GPT models typically consist of multiple Transformer [VSP*17] layers followed by a linear layer, which is referred to as the head layer. The transformer layers effectively encode a sequence into a latent embedding and the linear head is trained to perform a specific task, such as sequence generation or classification, using these latent values.

Numerous GPT-style models have been introduced, and among them, GPT-4 from Open AI [OA*23] has emerged as a top performer across various language-based tasks. However, GPT-4 is a closed-source solution. The leading open-source alternative is currently LLAMA2 [TMS*23], which surpasses other open-source LLMs in tasks related to commonsense reasoning, world knowledge, and reading comprehension.

LLMs have begun to garner attention in gesture-based tasks. For instance, Hensel et al. [HYT*23] uses ChatGPT [OA*23] for the selection and analysis of gestures, while Zeng et al. [ZWZ*23] uses

ChatGPT to analyse and comprehend performed gestures. To the best of our knowledge, there are no established methods for generating gestures directly from LLMs.

3. Method

In our exploration of using LLMs as a primary feature for co-speech gesture generation, we introduce LLAniMAtion. LLAniMAtion utilises LLAMA2 text embeddings, which can be used as an independent feature or in conjunction with PASE+ [RZP*20] audio features. We are one of the first to integrate a LLM in this way. The generative model is based on the adapted Transformer-XL architecture proposed by Windle et al. [WMMT23].

3.1. Speech Features

Our method can leverage both audio and text-based features. Each modality has differing sample rates, with audio sample values updating at a faster pace than text tokens. We extract features at their original sample rates and align them to fit the timing of a motion frame at 30fps. We use N to represent the number of $\approx$33ms motion frames in an input sequence. The PASE+ and LLAMA2 model weights are frozen and not updated during training.

3.1.1. Audio

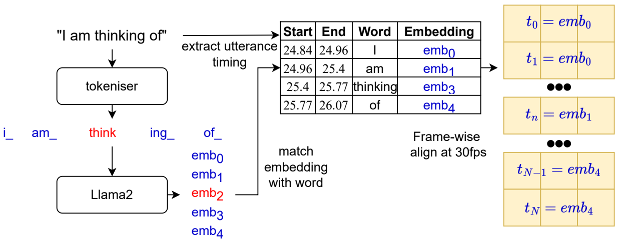
Audio features are extracted using the PASE+ model as these have been proven effective for gesture generation [WTGM22, WGT22, WMMT23]. PASE+ was trained by solving 12 regression tasks to learn important speech characteristics using a multi-task learning approach. These tasks include estimating MFCCs, FBANKs and other speech-related information, including prosody and speech content. Using this model, we extract audio feature embeddings of size 768 for each 33ms audio window to align with the 30fps motion. Consequently, audio feature vectors, A, with a shape of $(N, 768)$ are generated for each audio clip.

3.1.2. Text

Word-level features are extracted using the pre-trained, 7-billion parameter LLAMA2 model [TMS*23]. LLAMA2 adopts a Transformer architecture and has been trained on a corpus of 2 trillion tokens sourced from publicly available materials.

For each speech sequence, a transcript of the audio clip is tokenised and processed by the LLAMA2 model. We extract a sequence of embeddings using the transformer layers of the LLAMA2 model. The tokenised input is fed to the model and passed forward through all transformer layers but is not fed through any task-specific linear head. Therefore, an associated latent vector is extracted from the output of the last transformer layer for each word in the utterance and these are used as our text embedding. We perform a frame-wise alignment to ensure each embedding synchronises with its corresponding motion frame timing at 30fps. The process generates text-embedding vectors T of shape $(N, 4096)$.

Alignment is achieved by repeating text embeddings as needed to synchronise with the audio timing. This is an automated process using the transcript timings provided in the dataset; however, these timings could be extracted using automatic speech-to-text transcript methods such as OpenAI Whisper [RKX*23]. In instances where a word spans multiple frames, the vector is duplicated for the corresponding number of frames, and a zero-value vector is employed when no word is spoken at a specific frame. Figure 2 provides an overview of the alignment process.



Start	End	Word	Embedding
24.84	24.96	I	emb$_0$
24.96	25.4	am	emb$_1$
25.4	25.77	thinking	emb$_3$
25.77	26.07	of	emb$_4$

Figure 2: *Extracting text features using* LLAMA2*. The text is BPE-tokenised, and a* LLAMA2 *embedding is computed for each token. These embeddings are aligned with audio at 30fps by repeating frames as necessary.*

The input utterance is tokenised using a Byte-Pair Encoding (BPE) method, meaning a single word may be broken into multiple constituent parts. For example, the word “thinking” will be divided into two tokens, “think” and “ing”. In such cases, only the embedding for the last token is retained, and the embeddings for the preceding parts are discarded. For example, the embedding associated with “ing” is used rather than “think”. This is common practice when using LLMs as the final embedding is expected to encapsulate information about preceding tokens.

3.1.3. Speaker style

For each utterance, a speaker label is additionally provided as input. This is a unique ID per speaker which is passed through a learned embedding layer. The trainable weights of this layer ensure that speakers with similar gesture styles are positioned closely in the latent embedding space, while speakers with distinct gesturing styles are situated further apart. We use an 8-dimensional embedding to generate speaker vectors S with a shape of $(N, 8)$.

3.2. Body Pose Representation

The body pose at time n is defined as:

$$\mathbf{y_n} = [x_n, y_n, z_n, r_{j,1,n}, \ldots, r_{j,6,N}] \tag{1}$$

where x, y, z denote the global skeleton position and $r_{j,1:6,n}$ form rotations for each joint j in the 6D rotation representation presented by Zhou et al. [ZBL*19]. These values are standardised by subtracting the mean and dividing by the standard deviation computed from the training data.

3.3. Model Architecture

In this study, our primary objective is to evaluate the impact of LLM features on the animation of co-speech gestures. To accurately measure this effect, we employ an established model and

Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

training method. Specifically, we adopt a model based on the Cross-Attentive Transformer-XL, which demonstrated effectiveness in the Generation and Evaluation of Non-verbal Behaviour for Embodied Agent (GENEA) challenge 2023 [WMMT23]. This approach is built on the Transformer-XL model architecture [DYY*19] which uses segment-level recurrence with state reuse and a learned positional encoding scheme to ensure temporally cohesive boundaries between segments. Windle et al. extend this architecture using cross-attention to incorporate the second speaker's speech into the prediction when used in a dyadic setting. Notably, this architecture delivers high-quality results without the need for more involved training techniques such as diffusion.

Either a single modality or a combination of features are used to form the input feature vectors $X \in \{X_a, X_t, X_+, X_\times\}$. Please refer to Section 4.2 for more details on the construction of this matrix. We train our model on dyadic conversation between a main-agent and interlocutor. Specifically, we predict the main-agent's gesturing conditioned on both main-agent and interlocutor speech. Consequently, we compute a set of input features for each speaker, X^{ma} and X^{in}, and a set of target poses for the main-agent, Y. These extracted features are segmented into non-overlapping segments of length W frames.

Given an input feature vector X of length W, the Transformer-XL predicts $\hat{Y}$ of length W using a sliding window technique with no overlap. Consequently, for a speech sequence of length N, our model is invoked $\lceil \frac{N}{W} \rceil$ times. Figure 3 shows an overview of this approach.

3.4. Training Procedure

We follow the same training methodology as in Windle et al. [WMMT23] and include the same geometric and temporal constraints in the loss function. The loss function L_c comprises multiple terms including a L_1 loss on the rotations (L_r), positions (L_p), velocity (L_v), acceleration (L_a) and kinetic energy (L_{v^2}) of each joint.

All training parameters were kept the same as in Windle et al. However, an additional two self-attention layers were included in the Cross Attentive Transformer-XL. These additional layers were chosen based on validation loss values and the quality of the predicted validation sequences, as observed by our team. We train our models for 1300 epochs using the AdamW optimiser [LH17].

3.5. Smoothing

The raw model output can contain low levels of high-frequency noise. Following other work on motion synthesis [ZJGL23, ZBC22], we apply a Savitzky-Golay Smoothing filter to mitigate this. We use a window length of 9 and polynomial order of 2. The small window size and low polynomial means this filter provides a very small amount of localised smoothing while retaining accurate beat gestures.

4. Experimental Setup

Four distinct models are trained, each with a different set of features: 1) PASE+: An audio-only model, 2) LLAniMAtion: A LLAMA2 text-only model, 3) LLAniMAtion-+: A LLAMA2 and PASE+ concatenated model and 4) LLAniMAtion-×: A LLAMA2 and PASE+ cross-attention model. In this section we describe our data and details on the model configurations.

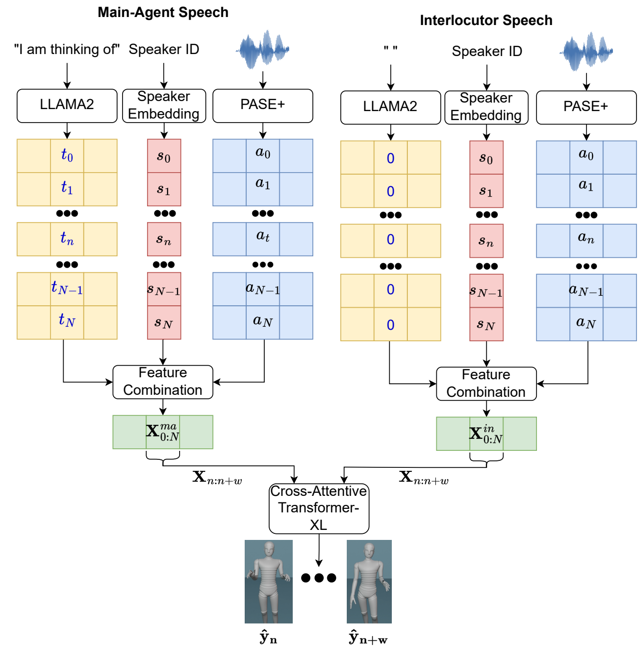


Figure 3: *Overview of LLAniMAtion method. Our model takes LLAMA2 features as input, along with a speaker embedding and optional PASE+ features that encode the speech of a main-agent and an interlocutor. The features are combined and processed through a cross-attentive Transformer-XL model that produces gesture animation for the main-agent.*

4.1. Data

The data used in this study is from the GENEA challenge 2023 [KNY*23]. This dataset is derived from the Talking With Hands dataset [LDM*19], containing dyadic conversations between a main-agent and interlocutor. It comprises high-quality 30fps motion capture data in Biovision Hierarchical (BVH) format. The dataset includes both speech audio and text transcripts derived from both speakers in the conversations.

The dataset is divided into three splits: 1) train, 2) validation, and 3) test. The validation set is employed for model tuning and refinement, while the test set is exclusively reserved for evaluation.

4.2. Feature Combinations

Our experiments use audio and text modalities in isolation and additionally investigate two approaches for combining the two modalities: 1) post-extraction concatenation and 2) cross-attention, respectively referred to as LLAniMAtion-+ and LLAniMAtion-×.

4.2.1. Single Modalities

To use each modality individually, we concatenate the speaker S matrices with the audio A or text T along the feature dimension to form X_a and X_t, respectively. The concatenated matrix is then passed through a linear layer, giving:

$$\begin{aligned} X_a &= W_a(A,S)^\top + \mathbf{b}_a \\ X_t &= W_t(T,S)^\top + \mathbf{b}_t \end{aligned} \qquad (2)$$

where W_a, W_t, $\mathbf{b}_a$ and $\mathbf{b}_t$ are learned parameters. X_a and X_t are used as inputs for training the single modality audio and text-based models respectively.

4.2.2. Concatenation

To combine modalities we concatenate A, T and S matrices along the feature dimension. The concatenated matrix is then passed through a linear layer, giving:

$$X_+ = W(A,T,S)^\top + \mathbf{b} \qquad (3)$$

where W and $\mathbf{b}$ are learned parameters. This results in X_+ which are the concatenated audio and text features and serve as the input to LLAniMAtion-$+$.

4.2.3. Cross-attention

Additionally, we experiment with using cross-attention for combining audio and text features. Cross-attention has been shown to be an effective method of combining modalities, as evidenced in Ng et al. [NRB*24]. In this approach, we first concatenate the style embedding to both audio and text features. We then linearly project the two concatenated matrices into the same feature dimension size, d, following Equation 2. We perform cross-attention on the feature dimension, such that the projected audio features, X_a, serve as the query, while the projected text features, X_t are set as the key and value [VSP*17]:

$$X_\times = \text{softmax}(\frac{X_a X_t^\top}{\sqrt{d}})X_t \qquad (4)$$

giving the cross attention combined audio and text features $X_\times$ which are used as input for training LLAniMAtion-$\times$.

5. Evaluation

We present an evaluation into the efficacy of LLAMA2 features for gesture generation, in isolation and in combination with audio PASE+ features. We present our observations and report the associated performance metrics. Finally, we describe a user study that measures the differences in perceived quality.

5.1. Observations

We observe noticeable differences between the animation produced by the PASE+-based model and the LLAniMAtion method. The PASE+ version primarily generates beat gestures, whereas the LLAniMAtion model exhibits more varied motions, encompassing both beat and semantic gestures. The animation from LLAniMAtion appears to be more expansive and confident. Video examples and comparisons showing this effect can be seen in the supplementary material.

5.1.1. Beat Gestures

Beat gestures are characterised by simple and fast movements of the hands, serving to emphasise prominent aspects of the speech [BP21]. These gestures have a close relationship with the timing of prosodic activity, such as acoustic energy and pitch [WTGM22, PHEGD20]. Given that these prosodic features can be directly derived from the audio signal, an audio-based model can be very effective at generating beat gestures. A beat gesture is not necessarily expected for every audio beat, but when performed, it is likely to be well-timed with the audio beats.

Using the motion and audio beat extraction method defined in the beat align score calculation proposed by Liu et al. [LZI*22], we can visualise the onset of audio beats and motion gestures over time. Remarkably, we observed that LLAniMAtion with LLAMA2 and no audio features consistently executes beat gestures in synchronisation with audio beats, despite lacking explicit energy or pitch information. Figure 4 shows a 1.5-second clip with the left wrist motion onsets in green and audio beat onsets in red. A speaker can be seen swiftly moving their left hand from left to right in time with audio beats and returning close to their original pose.

Although we temporally align the LLAMA2 embeddings providing the model with awareness of word timings, there is no explicit knowledge of syllable-level timing. Further investigation is needed; however, it is plausible that training with LLAMA2 embeddings may effectively encode information regarding the presence of lexically stressed syllables in context within words.

5.1.2. Semantic gestures

Semantic gestures are often directly linked to speech content, such as mimicking an action or nodding the head in agreement. In our observations, the LLAniMAtion method demonstrated superior performance compared to the audio PASE+-based model in generating these types of gestures.

In a test sequence where a speaker is describing the act of eating a crab, the LLAniMAtion gestures exhibit more activity compared to the PASE+ version, particularly when the speaker uses their hands to illustrate actions. This is exemplified when the hands mimicked sticking a fork in a crab for consumption in time with the verbal description. This sequence can be seen in Figure 1 and the supplementary video.

LLAniMAtion demonstrates the capacity to adequately encode agreeableness. For example, Figure 5 shows a predicted test sequence where the speaker can be seen nodding along with the word yes.

5.1.3. Laughter

During the transcription process of the GENEA dataset, laughter without speech was denoted using “###”. This representation was directly input to the LLAMA2 model for feature extraction. Although the generated embedding would not encode any semantic meaning, our model learns to associate these tokens to laughter. The LLAniMAtion method captures moments of laughter as illustrated in Figure 6, where the character partially creases over. This specific behaviour is not observed in the gesture animation produced by the PASE+-based model .

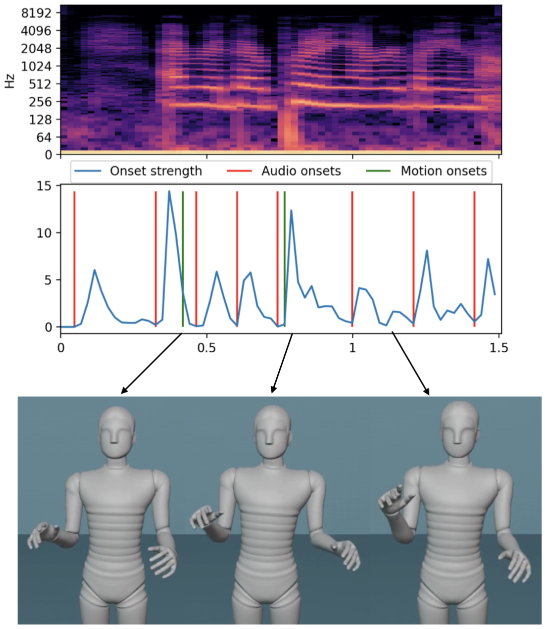


Figure 4: *Generated gestures for given audio beats using* LLAniMAtion *method. Using a 1.5s audio clip from the test dataset, we show the audio spectrogram, as well as aligned audio beat onsets and their corresponding onset strengths, as well as motion gesture onset detection of the left wrist using the method of beat detection defined in Liu et al. [LZI*22]. The speaker moves their left hand from right to left and back again as the syllables are stressed.*

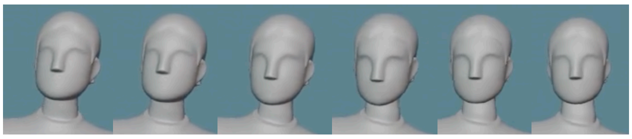


Figure 5: *Example nod motion temporally aligned with the word "yes" being spoken. from a test sequence generated using the* LLAniMAtion

5.2. Performance Metrics

Evaluating the objective performance of gesture generation poses a challenging research question, primarily due to the many-to-many ambiguous relationship between speech and gesture. No single metric has been developed that correlates with human perception. However a *combination* of metrics can be used as a means to somewhat evaluate the quality of generated gesture. Frèchet Gesture Distance (FGD) [YCL*20, BCRM21, NRB*24], Frèchet Kinetic Distance (FD$_k$) [NRB*24] and Beat Alignment (BA) [LYRK21, LZI*22] are useful metrics for this task. These metrics are indicative of static and dynamic appropriateness, and the alignment of motion to speech [ANBH23, LZI*22, YCL*20]. Frèchet Gesture Distance is a measure based on the Frèchet Inception Distance (FID) [HRU*17], which is commonly used for evaluating generative models. A pre-trained autoencoder extracts domain-specific latent features from both ground truth and predicted motion. The FGD score is a Frèchet distance between the two multivariate Gaussian distributions of these features in latent space. This measures similarity between the generated and ground truth poses but does not necessarily indicate how well the generated examples temporally align with the audio.

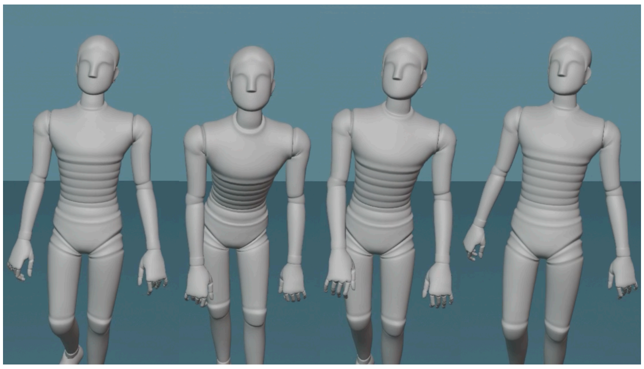


Figure 6: *Example laughter sequence generated using the* LLAniMAtion *method*

Model	*FGD* ↓	*FD$_k$* ↓	BA↑
PASE+	79.90	34.37	**0.871**
LLAniMAtion	61.86	24.23	0.855
LLAniMAtion-+	**47.56**	**23.79**	0.869
LLAniMAtion-×	66.87	25.70	0.865

Table 1: *Frèchet Gesture Distance (FGD) , Frèchet Kinetic Distance (FD$_k$) and Beat Alignment (BA) scores for each system calculated with respect to the ground truth test dataset.*

Frèchet Kinetic Distance is similar, however, there is no autoencoding process. Instead, the first derivative of each joint is used to determine the distribution of velocities for both the ground truth and predicted motion. FD$_k$ is the Frèchet Distance between these two distributions.

Beat Alignment has been adapted from music synthesis [LYRK21] to work with gesture generation [LZI*22]. Using a chamfer distance between audio and gesture beats, this gives a synchrony measure between the two. Beats are detected using the root mean square onset of the audio and a motion beat is identified by the local minimums of the velocity.

5.2.1. Results

The measures presented in Table 1 indicate that the FGD and FD$_k$ scores are consistently lower for all LLAMA2-based models than for the model trained on PASE+ features. This suggests that the motion generated by LLAniMAtion may be closer to ground truth, with LLAniMAtion-+ showing the most realistic motion. The BA score suggests that the audio features are the most timely, however, the differences between this and the LLAniMAtion methods are minimal. Notably, the method with no audio features is competitive in FGD and BA scores.

5.3. User Study

We present a user study to further evaluate perceived human likeness and appropriateness of the animations from the PASE+-based model compared with the LLAMA2-based LLAniMAtion method. Participants were hired through the Prolific platform with 50 participants in each experiment after removing any participants that failed attention checks. Participants were filtered to be fluent in English. For this study, we used a similar methodology to Alexanderson et al. [ANBH23] and the GENEA Challenge 2023 [KNY*23].

All test sequences for each method were rendered on the same virtual avatar released by Kucherenko et al. [KNY*23], as shown in Figure 1. We use the exact clip timings from the GENEA Challenge, comprising 41 clips with an average duration of 10 seconds each. During evaluation, users exclusively heard the audio of the main-agent being animated.

In our pairwise system comparison, participants were presented with two side-by-side videos generated for the same audio but with different systems. To mitigate bias, we randomise the question order and randomly swap the side of the screen that each condition is shown.

The question for all studies was posed as "Which character's motion do you prefer, taking into account both how natural-looking the motion is and how well it matches the speech rhythm and intonation?". The participants were asked to choose from the options {**Clear preference for *left***, **Slight preference for *left***, **No Preference**, **Slight preference for *right*** and **Clear preference for *right***}. The scoring methodology uses a merit system [PHS05] where an answer is given a value of 2, 1 or 0 for clear preference, slight preference and no preference, respectively. Preference testing allows a win rate calculation where a win is assigned when there is an identified preference for a system, not including ties. A one-way ANOVA test with a post-hoc Tukey test was subsequently used for significance testing.

5.3.1. Results

Table 2 summarises the results of the user study. These findings validate the objective measure scores in that all LLAniMAtion-based models outperform the PASE+ audio-only method. According to the merit score, all LLAniMAtion methods were significantly preferred over the PASE+ approach ($p < 0.001$). Win and tie rates show that LLAniMAtion methods win or are tied with PASE+ most of the time. Surprisingly, the highest win rate is recorded by the LLAniMAtion method with no PASE+ features included, suggesting that using text as a sole input is sufficient to generate plausible speech gesturing, and that audio features are somewhat redundant in our model

Between each LLAniMAtion method, there is no statistically significant difference in merit scores. We examine the win and tie rates against LLAniMAtion to determine if adding PASE+ features will provide additional preference. We can see from these rates that the choice between LLAniMAtion settings is almost tied to wins and losses. LLAniMAtion-$+$ wins 1.9% less than LLAniMAtion-$\times$; however, the tie rate is higher and therefore loses less than LLAniMAtion-$\times$.

This initial study concludes that LLAMA2 features are powerful at encoding information useful to gesture generation and can produce more realistic-looking gestures than a model trained on audio input. Combining modalities also does not make a significant difference, although the concatenation of features performs slightly better than the cross-attention regarding merit scores and win/tie rates.

6. Comparison Against Other Systems

Our previous study has shown that we achieve a significant performance improvement achieved by integrating LLM features into gesture-generation models. We now perform additional experiments to compare our best performing LLAniMAtion and LLAniMAtion-$+$ approaches against both ground truth and the current state-of-the-art method. This broader evaluation aims to assess performance across the field.

We compare against the state-of-the-art Contrastive Speech and Motion Pretraining Diffusion (CSMP-Diff) method [DMAB23], which achieved the highest human-likeness and speech appropriateness rating among the entries to the 2023 GENEA challenge.

Objective performance metrics are shown in Table 4. CSMP-Diff performs better in FGD and FD_k scores. We find minimal differences to the BA score, with LLAniMAtion marginally outperforming CSMP-Diff.

We repeated the user study following the protocol as described Section 5.3, and the results are summarised in Table 3. In terms of merit score, the ground truth was perceived as significantly better than any other method ($p < 0.001$), underscoring the current challenge in consistently generating human-realistic gesturing. CSMP-Diff was considered superior to both LLAniMAtion methods ($p < 0.001$). Despite this difference, when examining the win rates against CSMP-Diff, we find that the LLAniMAtion method wins 31.4% of the time and ties 16.1%. Meanwhile, our LLAniMAtion-$+$ method won 35.6% of the time and ties 14.4%. In each case, LLAniMAtion and LLAniMAtion-$+$ are rated as good or better than CSMP-Diff 47.5% and 50% of the time, respectively.

CSMP-Diff incorporates both diffusion and contrastive speech and motion pre-training, representing two advanced and complex techniques. Despite these sophisticated methods, our evidence indicates that LLAniMAtion, even in the absence of any audio input, can perform as well as or better than CSMP-Diff nearly half the time. This suggests that LLAMA2 features serve as incredibly valuable feature encodings for gesture animation. Simpler models that use more descriptive features may be beneficial, even if performance does not exceed the state-of-the-art but remains competitive. Training and inference times may be reduced, and the computational resources required may be lower when compared to more complicated methods like diffusion. While computational efficiency is not an aim of this paper, it is worth considering that using valuable, descriptive features such as LLAMA2 may benefit future efficiency savings.

7. Conclusion

We have explored the use of LLAMA2 features for speech-to-gesture generation in our proposed LLAniMAtion method. With the

		vs PASE+		vs LLAniMAtion		vs LLAniMAtion-+		vs LLAniMAtion-×	
	Merit Score	Win Rate	Tie Rate	Win Rate	Tie Rate	Win Rate	Tie Rate	Win Rate	Tie Rate
PASE+	0.37±0.05	-	-	25.4%	11.4%	24.6%	14.8%	28.4%	14.8%
LLAniMAtion	0.68±0.06	63.3%	11.4%	-	-	38.6%	22.3%	44.3%	14.8%
LLAniMAtion-+	**0.69**±0.06	61.0%	14.8%	39.0%	22.3%	-	-	43.2%	20.8%
LLAniMAtion-×	0.64±0.06	56.8%	14.8%	40.9%	14.8%	36.0%	20.8%	-	-

Table 2: *User study results comparing modality inclusion. Merit scores [PHS05] with 95% confidence intervals, win and tie rates for each comparison.*

		vs GT		vs LLAniMAtion		vs LLAniMAtion-+		vs CSMP-Diff	
	Merit Score	Win Rate	Tie Rate	Win Rate	Tie Rate	Win Rate	Tie Rate	Win Rate	Tie Rate
GT	1.16±0.05	-	-	78.6%	8.9%	74.8%	8.9%	68.9%	11.1%
LLAniMAtion	0.34±0.04	12.5%	8.9%	-	-	34.2%	33.6%	31.4%	16.1%
LLAniMAtion-+	0.36±0.04	16.4%	8.9%	32.2%	33.6%	-	-	35.6%	14.4%
CSMP-Diff	**0.58**±0.05	20%	11.1%	52.5%	16.1%	50.0%	14.4%	-	-

Table 3: *User study results comparing LLAniMAtion against CSMP-Diff and Ground Truth. Merit scores [PHS05] with 95% confidence intervals, win and tie rates for each comparison.*

Model	*FGD* ↓	FD_k ↓	BA↑
LLAniMAtion	61.86	24.23	0.855
LLAniMAtion-+	47.56	23.79	**0.869**
CSMP-Diff	**30.620**	**12.61**	0.866

Table 4: *Frèchet Gesture Distance (FGD) [YCL*20], Frèchet Kinetic Distance (FD_k) and Beat Alignment (BA) [LZI*22] scores for each system calculated with respect to the ground truth test dataset.*

use of LLAMA2 features we were able to generate well timed and contextually rich gestures even without the inclusion of any audio feature embedding. We explored the use of combining both audio and text modalities through concatenation and cross-attention and found that there was no significant difference in the inclusion of PASE+ features when compared to using LLAMA2 features in isolation. We have demonstrated the performance improvements when incorporating the LLAMA2 features into a gesture-generation model through both objective and subjective measures. Given this finding, we believe that human speech related gesture animation is heavily related to the semantic encoding that is present in the LLAMA2 embeddings, and that these embeddings additionally capture a notion of prosody from the language context. This is a somewhat surprising finding, and a result we think can have great practical impact on future content-aware animation systems.

We additionally compared our LLAniMAtion approach to ground truth as well as the state-of-the-art CSMP-Diff approach. The evaluation revealed that both LLAniMAtion and CSMP-Diff have areas where improvement is possible as they are unmatched against ground truth. While CSMP-Diff remains state-of-the-art, it is a complex model and our simpler alternative was rated as good or better than it 50% of the time. We predict that integrating LLM features into state-of-the-art systems will be a step towards bridging the gap between machine-generated and natural gesturing.

7.1. Future Work

While we show that the use of LLM features can be powerful for generating contextually and semantically correct gestures, more work is required to get performance closer to the ground truth. We use the 7-billion parameter release of LLAMA2 in this work due to hardware constraints. With more resources, the larger 70-billion parameter could be utilised, which may produce more nuanced and varied gesturing. We do not fine-tune the LLAMA2 model for our domain. LLMs are known to perform well with prompting and in-context learning [YJT*23] to fine-tune the model. We therefore foresee many opportunities for further performance gain.

This experiment explores speech during a natural, dyadic conversation. This is not scripted or acted for a particular affective state where the emotion of the speech may alter how the speaker moves. Human communication has nuance regarding the tone of someone's voice, influencing how an utterance should be interpreted, which may also affect the speaker's movements. The dataset used here does not include particular affective state labelling or an exceptionally diverse set of affective states; however, more work could be done to explore how well LLM features may handle these nuanced moments where speaker tone may modify the meaning of a phrase. We predict this may be where including both prosodic and semantic features may be the optimal solution.

References

[AGL*22] Ao T., Gao Q., Lou Y., Chen B., Liu L.: Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics (TOG) 41*, 6 (2022), 1–19. 2

[AHKB20] Alexanderson S., Henter G. E., Kucherenko T., Beskow J.: Style-controllable speech-driven gesture synthesis using normalising flows. In *Computer Graphics Forum* (2020), vol. 39, Wiley Online Library, pp. 487–496. 2

[ANBH23] Alexanderson S., Nagy R., Beskow J., Henter G. E.: Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Trans. Graph. 42*, 4 (2023), 1–20. doi:10.1145/3592458. 2, 6, 7

[BCRM21] Bhattacharya U., Childs E., Rewkowski N., Manocha D.: Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In *Proceedings of the 29th ACM International Conference on Multimedia* (2021), pp. 2027–2036. 6

[BGJM17] Bojanowski P., Grave E., Joulin A., Mikolov T.: Enriching word vectors with subword information. *Transactions of the association for computational linguistics 5* (2017), 135–146. 2

[BP21] Bosker H. R., Peeters D.: Beat gestures influence which speech sounds you hear. *Proceedings of the Royal Society B 288*, 1943 (2021), 20202419. 1, 5

[BRB*21] Bhattacharya U., Rewkowski N., Banerjee A., Guhan P., Bera A., Manocha D.: Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE virtual reality and 3D user interfaces (VR)* (2021), IEEE, pp. 1–10. 2

[BZMA20] Baevski A., Zhou Y., Mohamed A., Auli M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems 33* (2020), 12449–12460. 2

[DCLT18] Devlin J., Chang M.-W., Lee K., Toutanova K.: Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018). 2

[DMAB23] Deichler A., Mehta S., Alexanderson S., Beskow J.: Diffusion-based co-speech gesture generation using joint text and audio representation. In *Proceedings of the 25th International Conference on Multimodal Interaction* (2023), pp. 755–762. 7

[DRBD12] De Ruiter J. P., Bangerter A., Dings P.: The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Topics in Cognitive Science 4*, 2 (2012), 232–248. 1

[DYY*19] Dai Z., Yang Z., Yang Y., Carbonell J., Le Q. V., Salakhutdinov R.: Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860* (2019). 4

[HES*22] Habibie I., Elgharib M., Sarkar K., Abdullah A., Nyatsanga S., Neff M., Theobalt C.: A motion matching-based framework for controllable gesture synthesis from speech. In *ACM SIGGRAPH 2022 Conference Proceedings* (2022), pp. 1–9. 2

[HKS*18] Hasegawa D., Kaneko N., Shirakawa S., Sakuta H., Sumi K.: Evaluation of speech-to-gesture generation using bi-directional lstm network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (2018), pp. 79–86. 2

[HRU*17] Heusel M., Ramsauer H., Unterthiner T., Nessler B., Hochreiter S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems 30* (2017). 6

[HYT*23] Hensel L. B., Yongsatianchot N., Torshizi P., Minucci E., Marsella S.: Large language models in textual analysis for gesture selection. In *Proceedings of the 25th International Conference on Multimodal Interaction* (2023), pp. 378–387. 2

[Ken94] Kendon A.: Do gestures communicate? a review. *Research on language and social interaction 27*, 3 (1994), 175–200. 1

[KHH*19] Kucherenko T., Hasegawa D., Henter G. E., Kaneko N., Kjellström H.: Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (2019), pp. 97–104. 2

[KNN*21] Kucherenko T., Nagy R., Neff M., Kjellström H., Henter G. E.: Multimodal analysis of the predictability of hand-gesture properties. *arXiv preprint arXiv:2108.05762* (2021). 2

[KNY*23] Kucherenko T., Nagy R., Yoon Y., Woo J., Nikolov T., Tsakov M., Henter G. E.: 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* (2023), pp. 792–801. 4, 7

[LDM*19] Lee G., Deng Z., Ma S., Shiratori T., Srinivasa S. S., Sheikh Y.: Talking with hands 16.2 m: 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* (2019), pp. 763–772. 4

[LH17] Loshchilov I., Hutter F.: Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017). 4

[LYRK21] Li R., Yang S., Ross D. A., Kanazawa A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 13401–13412. 6

[LZI*22] Liu H., Zhu Z., Iwamoto N., Peng Y., Li Z., Zhou Y., Bozkurt E., Zheng B.: Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European Conference on Computer Vision* (2022), Springer, pp. 612–630. 5, 6, 8

[McN85] McNeill D.: So you think gestures are nonverbal? *Psychological review 92*, 3 (1985), 350. 1

[NKA*23] Nyatsanga S., Kucherenko T., Ahuja C., Henter G. E., Neff M.: A comprehensive review of data-driven co-speech gesture generation. *Computer Graphics Forum 42*, 2 (2023), 569–596. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14776, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14776, doi:https://doi.org/10.1111/cgf.14776. 2

[NRB*24] Ng E., Romero J., Bagautdinov T., Bai S., Darrell T., Kanazawa A., Richard A.: From audio to photoreal embodiment: Synthesizing humans in conversations. *arXiv preprint arXiv:2401.01885* (2024). 2, 5, 6

[OA*23] OpenAI, :, Achiam J., Adler S., Agarwal S., Ahmad L., Akkaya I., Aleman F. L., Almeida D., Altenschmidt J., Altman S., Anadkat S., Avila R., Babuschkin I., Balaji S., Balcom V., Baltescu P., Bao H., Bavarian M., Belgum J., Bello I., Berdine J., Bernadett-Shapiro G., Berner C., Bogdonoff L., Boiko O., Boyd M., Brakman A.-L., Brockman G., Brooks T., Brundage M., Button K., Cai T., Campbell R., Cann A., Carey B., Carlson C., Carmichael R., Chan B., Chang C., Chantzis F., Chen D., Chen S., Chen R., Chen J., Chen M., Chess B., Cho C., Chu C., Chung H. W., Cummings D., Currier J., Dai Y., Decareaux C., Degry T., Deutsch N., Deville D., Dhar A., Dohan D., Dowling S., Dunning S., Ecoffet A., Eleti A., Eloundou T., Farhi D., Fedus L., Felix N., Fishman S. P., Forte J., Fulford I., Gao L., Georges E., Gibson C., Goel V., Gogineni T., Goh G., Gontijo-Lopes R., Gordon J., Grafstein M., Gray S., Greene R., Gross J., Gu S. S., Guo Y., Hallacy C., Han J., Harris J., He Y., Heaton M., Heidecke J., Hesse C., Hickey A., Hickey W., Hoeschele P., Houghton B., Hsu K., Hu S., Hu X., Huizinga J., Jain S., Jain S., Jang J., Jiang A., Jiang R., Jin H., Jin D., Jomoto S., Jonn B., Jun H., Kaftan T., Łukasz Kaiser, Kamali A., Kanitscheider I., Keskar N. S., Khan T., Kilpatrick L., Kim J. W., Kim C., Kim Y., Kirchner H., Kiros J., Knight M., Kokotajlo D., Łukasz Kondraciuk, Kondrich A., Konstantinidis A., Kosic K., Krueger G., Kuo V., Lampe M., Lan I., Lee T., Leike J., Leung J., Levy D., Li C. M., Lim R., Lin M., Lin S., Litwin M., Lopez T., Lowe R., Lue P., Makanju A., Malfacini K., Manning S., Markov T., Markovski Y., Martin B., Mayer K., Mayne A., McGrew B., McKinney S. M., McLeavey C., McMillan P., McNeil J., Medina D., Mehta A., Menick J., Metz L., Mishchenko A., Mishkin P., Monaco V., Morikawa E., Mossing D., Mu T., Murati M., Murk O., Mély D., Nair A., Nakano R., Nayak R., Neelakantan A., Ngo R., Noh H., Ouyang L., O'Keefe C., Pachocki J., Paino A., Palermo J., Pantuliano A., Parascandolo G., Parish J., Parparita E., Passos A., Pavlov M., Peng A., Perelman A., de Avila Belbute Peres F., Petrov M., de Oliveira Pinto H. P., Michael, Pokorny, Pokrass M., Pong V., Powell T., Power A., Power B., Proehl E., Puri R., Radford A., Rae J., Ramesh A., Raymond C., Real F., Rimbach K., Ross C., Rotsted B., Roussez H., Ryder N., Saltarelli M., Sanders T., Santurkar S., Sastry G., Schmidt H., Schnurr D., Schulman J., Selsam D., Sheppard K., Sherbakov T., Shieh J., Shoker S., Shyam P., Sidor S., Sigler E., Simens M., Sitkin J., Slama K., Sohl I., Sokolowsky B., Song Y., Staudacher N., Such F. P., Summers N., Sutskever I., Tang J., Tezak N., Thompson M., Tillet P., Tootoonchian A., Tseng E., Tuggle P., Turley N., Tworek J., Uribe J. F. C., Vallone A., Vijayvergiya A., Voss C., Wainwright C., Wang J. J., Wang A., Wang B., Ward J., Wei J., Weinmann C., Welihinda A., Welinder P., Weng J., Weng L., Wiethoff M., Willner D., Winter C., Wolrich S., Wong H., Workman L., Wu S., Wu J., Wu M., Xiao K., Xu T., Yoo S., Yu K., Yuan Q., Zaremba W., Zellers R., Zhang C., Zhang M., Zhao S., Zheng T., Zhuang J., Zhuk W., Zoph B.: Gpt-4 technical report, 2023. arXiv:2303.08774. 2

[PHEGD20] Pouw W., Harrison S. J., Esteve-Gibert N., Dixon J. A.: Energy flows in gesture-speech physics: The respiratory-vocal system and its coupling with hand gestures. *The Journal of the Acoustical Society of America 148*, 3 (2020), 1231–1247. 2, 5

[PHS05] Parizet E., Hamzaoui N., Sabatie G.: Comparison of some listening test methods: a case study. *Acta Acustica united with Acustica 91*, 2 (2005), 356–364. 7, 8

[PKJS20] Pang K., Komura T., Joo H., Shiratori T.: Cgvu: Semantics-guided 3d body gesture synthesis. In *Proc. GENEA Workshop. https://doi. org/10.5281/zenodo* (2020), vol. 4090879. 2

[PSM14] Pennington J., Socher R., Manning C. D.: Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (2014), pp. 1532–1543. 2

[QTZ*21] Qian S., Tu Z., Zhi Y., Liu W., Gao S.: Speech drives templates: Co-speech gesture synthesis with learned templates. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 11077–11086. 2

[RKX*23] Radford A., Kim J. W., Xu T., Brockman G., McLeavey C., Sutskever I.: Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning* (2023), PMLR, pp. 28492–28518. 3

[RZP*20] Ravanelli M., Zhong J., Pascual S., Swietojanski P., Monteiro J., Trmal J., Bengio Y.: Multi-task self-supervised learning for robust speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020), IEEE, pp. 6989–6993. 2, 3

[SK94] Studdert-Kennedy M.: Hand and mind: What gestures reveal about thought. *Language and Speech 37*, 2 (1994), 203–209. 1

[TMS*23] Touvron H., Martin L., Stone K., Albert P., Almahairi A., Babaei Y., Bashlykov N., Batra S., Bhargava P., Bhosale S., et al.: Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023). 2, 3

[VSP*17] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I.: Attention is all you need. *Advances in neural information processing systems 30* (2017). 2, 5

[WGT22] Windle J., Greenwood D., Taylor S.: Uea digital humans entry to the genea challenge 2022. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (2022), pp. 771–777. 2, 3

[WMMT23] Windle J., Matthews I., Milner B., Taylor S.: The uea digital humans entry to the genea challenge 2023. In *Proceedings of the 25th International Conference on Multimodal Interaction* (New York, NY, USA, 2023), ICMI '23, Association for Computing Machinery, p. 802–810. URL: https://doi.org/10.1145/3577190.3616116, doi:10.1145/3577190.3616116. 2, 3, 4

[wor13] word2vec, 2013. Accessed on Jan 2024. URL: https://code.google.com/archive/p/word2vec/. 2

[WTGM22] Windle J., Taylor S., Greenwood D., Matthews I.: Arm motion symmetry in conversation. *Speech Communication 144* (2022), 75–88. 2, 3, 5

[YCL*20] Yoon Y., Cha B., Lee J.-H., Jang M., Lee J., Kim J., Lee G.: Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG) 39*, 6 (2020), 1–16. 2, 6, 8

[YJT*23] Yang J., Jin H., Tang R., Han X., Feng Q., Jiang H., Yin B., Hu X.: Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712* (2023). 2, 8

[YKJ*19] Yoon Y., Ko W.-R., Jang M., Lee J., Kim J., Lee G.: 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)* (2019), IEEE, pp. 4303–4309. 2

[ZBC22] Zhou C., Bian T., Chen K.: Gesturemaster: Graph-based speech-driven gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (2022), pp. 764–770. 2, 4

[ZBL*19] Zhou Y., Barnes C., Lu J., Yang J., Li H.: On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019). 3

[ZJGL23] Zhang F., Ji N., Gao F., Li Y.: Diffmotion: Speech-driven gesture synthesis using denoising diffusion model. In *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part I* (2023), Springer, pp. 231–242. 4

[ZWZ*23] Zeng X., Wang X., Zhang T., Yu C., Zhao S., Chen Y.: Gesturegpt: Zero-shot interactive gesture understanding and grounding with large language model agents. *arXiv preprint arXiv:2310.12821* (2023). 2