

DeepGesture: A conversational gesture synthesis system based on emotions and semantics

Thanh Hoang-Minh

hmthanhhgm@gmail.com

OpenHuman.ai

Ho Chi Minh City, Vietnam

Department of Information Technology, VNUHCM - University of Science

Ho Chi Minh City, Vietnam

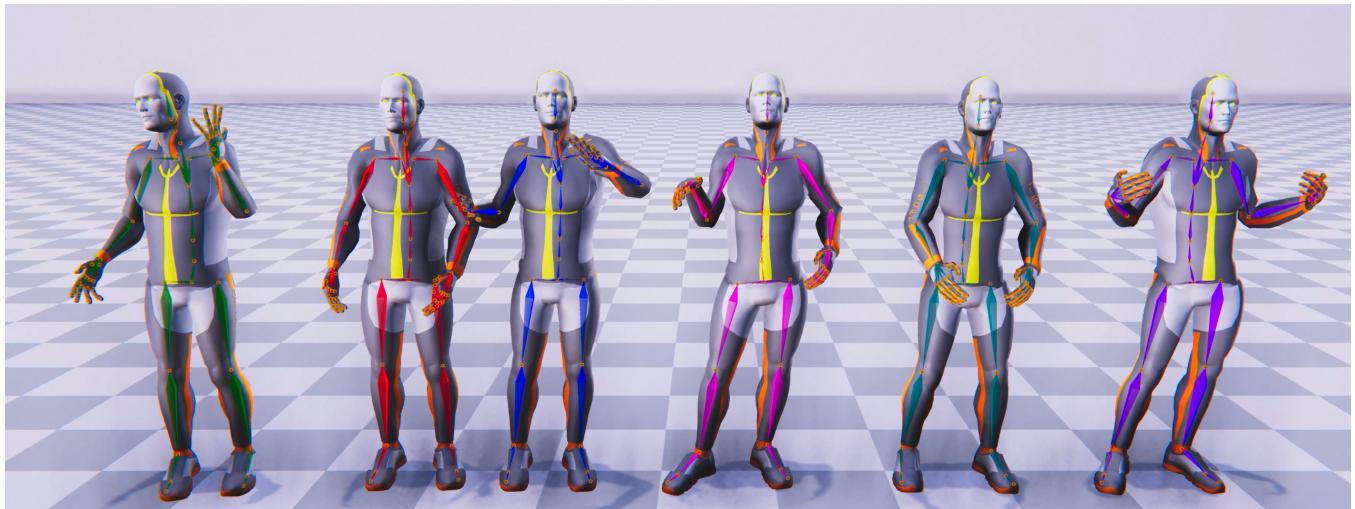


Figure 1: Gesture generation result in various emotion, speech and text

Abstract

Along with the explosion of large language models, improvements in speech synthesis, advancements in hardware, and the evolution of computer graphics, the current bottleneck in creating digital humans lies in generating character movements that correspond naturally to text or speech inputs.

In this work, we present **DeepGesture**, a diffusion-based gesture synthesis framework for generating expressive co-speech gestures conditioned on multimodal signals—text, speech, emotion, and seed motion. Built upon the DiffuseStyleGesture model, DeepGesture introduces novel architectural enhancements that improve semantic alignment and emotional expressiveness in generated gestures. Specifically, we integrate fast text transcriptions as semantic conditioning and implement emotion-guided classifier-free diffusion to support controllable gesture generation across affective states. A lightweight Transformer backbone combines full self-attention

and cross-local attention for effective feature fusion of heterogeneous modalities. To visualize results, we implement a full rendering pipeline in Unity based on BVH output from the model. Evaluation on the ZeroEGGS dataset shows that DeepGesture produces gestures with improved human-likeness and contextual appropriateness, outperforming baselines on Mean Opinion Score and Fréchet Gesture Distance metrics. Our system supports interpolation between emotional states and demonstrates generalization to out-of-distribution speech, including synthetic voices—marking a step forward toward fully multimodal, emotionally aware digital humans.

CCS Concepts

• Computing methodologies → Animation; Natural language processing; Neural networks.

Keywords

co-speech gesture synthesis, gesture generation, character animation, diffusion model, neural generative model

ACM Reference Format:

Thanh Hoang-Minh. 2026. DeepGesture: A conversational gesture synthesis system based on emotions and semantics. In *Proceedings of ACM SIGGRAPH Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '26)*. ACM, New York, NY, USA, 18 pages. <https://doi.org/10.1145/XXXXXXX.XXXXXXX>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGGRAPH '26, Los Angeles, CA, USA

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/26/07

<https://doi.org/10.1145/XXXXXXX.XXXXXXX>

1 INTRODUCTION

Every day, billions of people around the world look at RGB screens, and the output displayed on these screens is the result of various software systems. Therefore, the rendering of each pixel on the screen and the realistic simulation of images have been a focus of computer graphics scientists since the 1960s, particularly in the simulation of human figures or digital human.

Today, computer graphics technology can realistically simulate many complex objects such as water, roads, bread, and even human bodies and faces with incredible detail, down to individual hair strands, pimples, and eye textures. In 2015, using 3D scanning techniques [Metallo et al. 2015] to capture all angles of the face and light reflection, researchers were able to recreate President Obama's face on a computer with high precision, making it almost indistinguishable from the real thing.

Artificial intelligence (AI) has shown remarkable results in recent years, not only in research but also in practical applications, such as ChatGPT and Midjourney, showcasing vertical and horizontal growth in various fields. Although computer graphics can construct highly realistic human faces, gesture generation has traditionally relied on Motion Capture from sensors, posing significant challenges in building an AI system that learns from data. Generating realistic beat gestures is challenging because gestural beats and verbal stresses are not strictly synchronized, and it is complicated for end-to-end learning models to capture the complex relationship between speech and gestures.

With the success of large language models in text processing and the advancement of Computer-generated imagery (CGI) in producing nearly indistinguishable human faces, combined with the increasing ease and accuracy of human speech synthesis, gesture generation through AI has become one of the main bottlenecks in developing interactive digital human.

1.1 Problem Data

1.1.1 Skeleton Structure of Gestures. A gesture is defined as the movement of a character's entire body over time, as shown in Figure 2, captured frame by frame. In computer graphics, character motion is represented as bone-specific movements, including hands, legs, head, spine, etc. The full character skeleton structure is presented in Appendix Subsection A.1.

Motion data is captured using motion capture systems using cameras and specialized sensors. The output is typically stored in BVH (Biovision Hierarchy) files.

A BVH file consists of two main parts: HIERARCHY and MOTION. The HIERARCHY section is structured as a tree containing the skeleton's initial positions and names. The MOTION section contains the movement data for the entire skeleton frame-by-frame. Each BVH file includes frame rate (fps) and total frame count. Details are presented in Appendix Subsection A.2.

1.1.2 Motion Structure of Gestures. Gesture motion data, as illustrated in Figure 2, or the MOTION section of a BVH file, contains position and rotation information per frame. Each frame is a skeleton of 75 bones: $\{b_1, b_2, \dots, b_{75}\}$, where each bone has position $\{p_x, p_y, p_z\}$ and rotation $\{r_x, r_y, r_z\}$.

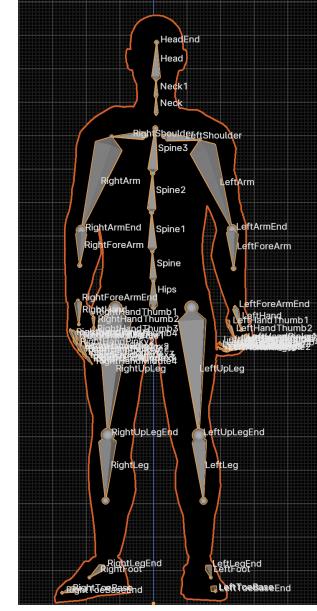


Figure 2: Skeleton and joint names of single frame

The output of gesture generation is a sequence of bone rotations per frame. The generated gestures are evaluated based on naturalness, human-likeness, and contextual appropriateness.

The skeleton's position and rotation data are preprocessed into a feature vector $g \in \mathbb{R}^D$ with $D = 1141$. The learning data becomes $x \in \mathbb{R}^{M \times D}$. The preprocessing pipeline is detailed in Appendix A.

1.2 Problem Statement

The ultimate goal is to produce a sequence of gestures that reflect the motion of the skeleton frame by frame. This can be approached via classification, clustering, or regression. Gesture generation is approached in this work as a regression-based prediction problem, wherein the next sequence is generated conditioned on the current gesture input.

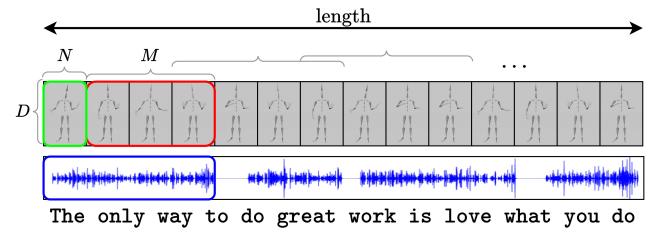


Figure 3: A gesture sequence: the first N frames are used as seed gesture s , and the remaining M frames are to be predicted

Each gesture sequence is labeled with an emotion. A key novelty of our approach is pairing the gesture sequence with both the original speech and the corresponding text (transcribed from the speech).

The objective is to build a model that predicts M future frames from the given inputs: seed gesture $\mathbf{s} \in \mathbb{R}^{1:N \times D}$, speech \mathbf{a} , text \mathbf{v} , and emotion \mathbf{e} .

The model prediction is $\hat{\mathbf{x}} \in \mathbb{R}^{1:M \times D}$, which is compared against ground-truth gesture $\mathbf{x} \in \mathbb{R}^{1:M \times D}$.

Input	Predicted Output	Ground Truth
<ul style="list-style-type: none"> Seed gesture sequence: $\mathbf{s} \in \mathbb{R}^{1:N \times D}$ Speech signal: \mathbf{a} Text: \mathbf{v} Emotion: \mathbf{e} (Happy, Sad, Neutral, Angry, Old, Funny) 	<ul style="list-style-type: none"> Predicted gesture sequence: $\hat{\mathbf{x}} \in \mathbb{R}^{1:M \times D}$ 	<ul style="list-style-type: none"> Ground truth gesture: $\mathbf{x} \in \mathbb{R}^{1:M \times D}$

1.3 Challenges

There are several challenges in building a model that can learn human-like conversational gesture patterns:

- (1) *Limited and low-quality data*: Creating large-scale, high-quality datasets for motion capture is extremely costly in the industry.
- (2) *Inconsistent context between modalities*: Text datasets are more abundant than speech, and speaker attribution is often missing. Synchronization between speech and emotional tone is also lacking. Additionally, training texts span many unrelated topics.
- (3) *Imbalanced feature distributions*: Current datasets are biased toward English-speaking gestures, with imbalanced gesture distributions between speaking, questioning, and silent states.
- (4) *High computational cost due to multimodal input*: The model must encode text, speech, and 3D pose data, increasing the computational load during both training and inference. Reducing input information also degrades performance.
- (5) *Sequential preprocessing steps*: Although human-computer interaction is most effective through speech and keyboard input, processing the text and speech input for gesture generation must be done sequentially. In real-world applications, inference latency is critical, and users cannot wait long. Rendering the gestures on screen must also be optimized for speed.

Contributions

In summary, our main contributions in this paper are:

- From existing datasets, speech is transcribed into text and used as additional semantic features for training.
- Based on the DiffuseStyleGesture model, we extend the conditional denoising process to include text features.
- In this work, we use Unity for rendering, data extraction, and visualization of gesture generation results.
- A rendering pipeline is designed and implemented in this paper, with the system demonstrated using Unity.

2 RELATED WORK

Gesture generation, like other machine learning tasks, has been explored using both traditional rule-based methods and modern data-driven approaches. We first examine the fundamental characteristics of gestures (Subsection 2.1) as a basis for modeling their relationship with text and speech. Figure 5 illustrates the common processing stages shared across various gesture generation techniques.

2.1 Gesture Characteristics

According to linguistics, gestures can be categorized into six main groups: adaptors, emblems, deictics, iconics, metaphorics, and beats [Ekman and Friesen 1969], [Sebeok and Umiker-Sebeok 2011]. Among them, beat gestures do not carry direct semantic meaning but play an important role in synchronizing rhythm between speech and gesture [Kipp 2005], [Sebeok and Umiker-Sebeok 2011]. However, the rhythm between speech and beat gestures is not fully synchronized, making the temporal alignment between them complex to model [McClave 1994], [Bhattacharya et al. 2021], [Kucherenko et al. 2020], [Yoon et al. 2020].

Gestures interact with various levels of information in speech [Sebeok and Umiker-Sebeok 2011]. For instance, emblematic gestures like a thumbs-up are usually linked to high-level semantic content (e.g., “good” or “excellent”), while beat gestures often accompany low-level prosodic features such as emphasis. Previous studies typically extract features from the final layer of the speech encoder to synthesize gestures [Alexanderson et al. 2020], [Bhattacharya et al. 2021], [Kucherenko et al. 2021], [Qian et al. 2021], [Yoon et al. 2022]. However, this approach may blend information from different levels, making it hard to disentangle rhythm and semantics.

As shown in linguistic studies [Kipp 2005], [Neff et al. 2008], [Webb 1997], gestures in daily communication can be decomposed into a limited number of semantic units with various motion variations. Based on this assumption, speech features are divided into two types: high-level features representing semantic units and low-level features capturing motion variations. Their relationships are learned across different layers of the speech encoder. Experiments demonstrate that this mechanism can effectively disentangle features at different levels and generate gestures that are both semantically and stylistically appropriate.

2.2 Overview of Gesture Generation Methods

2.2.1 Rule-Based Methods. These methods rely on clearly defined rules, which are manually crafted to determine how the system processes inputs to produce outputs.

Methods: Representative rule-based methods include the *Robot behavior toolkit* [Huang and Mutlu 2012] and *Animated*

conversation [Cassell et al. 1994]. These approaches typically map speech to gesture units using handcrafted rules. Rule-based systems allow for straightforward control over model outputs and provide good interpretability. However, the cost of manually designing these rules is prohibitive for complex applications requiring the processing of large-scale data.

2.2.2 Statistical Methods. These methods rely on data analysis, learning patterns from datasets, and using probabilistic models or mathematical functions for prediction. The approach involves optimizing model parameters to fit the data.

Methods: Like rule-based methods, data-driven methods also map speech features to corresponding gestures. However, instead of manual rules, they employ automatic learning based on statistical data analysis.

Representative statistical approaches include *Gesture controllers* [Levine et al. 2010], *Statistics-based* [Yang et al. 2020], which use probabilistic distributions to find similarities between speech and gesture features. *Gesture modeling* [Neff et al. 2008] constructs probabilistic models to learn individual speaker styles.

2.2.3 Deep Learning Methods. These methods utilize multi-layer perceptrons (MLPs) to automatically extract features from raw data and learn complex data representations through parameter optimization.

As illustrated in Figure 4, deep learning-based gesture generation methods can be divided into two main groups: likelihood-based models and implicit generative models [Song 2021].

Likelihood-Based Models

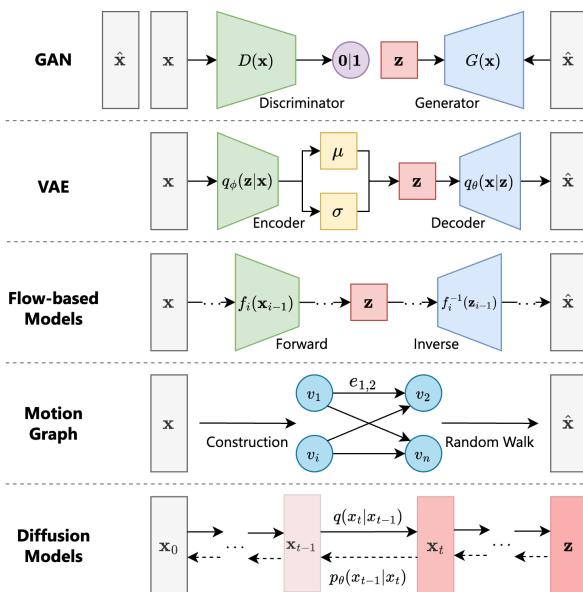


Figure 4: Overview of different generative models.

These models work by maximizing the likelihood of observed data given model parameters θ . The objective is to find the optimal parameters θ' by modeling the probability $p(\mathbf{x})$ of the data, where \mathbf{x} represents the gesture sequence.

Methods: The application of deep learning to gesture generation has evolved alongside the development of deep learning models, including RNNs, LSTMs, and Transformers. Representative likelihood-based methods include:

- *Gesticulator* [Kucherenko et al. 2020], which uses a Multilayer Perceptron (MLP) to encode text and audio features, with BERT-derived vectors used as text features.
- *HA2G* [Liu et al. 2022a] builds a hierarchical Transformer-based model to learn multi-level correlations between speech and gestures, from local to global features.
- *Gesture Generation from Trimodal Context* [Yoon et al. 2020] uses an RNN architecture and treats gesture generation as a translation task in natural language processing.
- *DNN* [Chiu et al. 2015] combines LSTM and GRU to build a classifier neural network that selects appropriate gesture units based on speech input.
- *Cascaded Motion Network (CaMN)* [Liu et al. 2022b] introduces the BEAT dataset and a waterfall-like model. Speaker information, emotion labels, text, speech, and gesture features are processed through layers to extract latent vectors. In the fusion stage, CaMN combines features sequentially: speaker and emotion features are merged first, followed by integration with latent vectors of text, speech, and gestures.
- **Motion Graph:** In *Gesturmaster* [Zhou et al. 2022], a semantic graph is constructed where words in a sentence are connected based on semantic relationships. The model then selects the most relevant nodes and edges in the graph to represent gestures.

Implicit Generative Models

Implicit generative models learn the data distribution without explicitly modeling the probability density function $p(\mathbf{x})$. Instead of directly computing $p(\mathbf{x})$, the model learns a mapping $G_\theta : \mathcal{Z} \rightarrow \mathcal{X}$ by matching the distributions between real and generated data $\mathbf{x} = G_\theta(\mathbf{z})$, $\mathbf{z} \sim p_z(\mathbf{z})$. Here, \mathcal{Z} denotes the input noise space, typically with a simple distribution p_z (e.g., Gaussian, Uniform), and \mathcal{X} is the space of real data, which in this case is gesture sequences \mathbf{x} .

In gesture generation, to incorporate conditions such as speech or text labels, implicit generative models often introduce a condition \mathbf{c} representing the context of the task, resulting in the conditional generation: $\mathbf{x} = G_\theta(\mathbf{z}, \mathbf{c})$. This context may include speech, text, speaker ID, style, emotion, or initial gestures.

A typical example is the generative adversarial network (GAN) and Diffusion model, where data is synthesized by

transforming an initial simple distribution (e.g., Gaussian) into the target data distribution.

Methods: Representative implicit generative models include:

- **MoGlow** [Henter et al. 2020] uses Normalizing Flows to maintain motion consistency and compatibility, while providing control via input parameters. This allows users to generate new motions or modify existing ones by adjusting control parameters.
- **GAN**: *GRU-based WGAN* [Wu et al. 2021] utilizes Wasserstein GAN to evaluate and improve the quality of gesture synthesis. The model focuses on optimizing the Wasserstein loss, mitigating mode collapse typically seen in traditional GANs. GRUs process speech data and convert it into usable gesture features, which are then fed into the WGAN for evaluation and refinement.
- **VAE**: *FreeMo* [Xu et al. 2022] employs a VAE to decompose gestures into pose modes and rhythmic motions. Gesture poses are randomly sampled using conditional sampling in the VAE’s latent space, while rhythmic motions are generated from...
- **VQ-VAE**: *Rhythmic Gesticulator* [Ao et al. 2022] pre-processes speech segments based on beats, dividing speech into smaller parts and representing them as blocks with normalized dimensions. Similar to the VQ-VAE approach, normalized gesture sequences are quantized into discrete gesture lexicons. The model learns the gesture vocabulary conditioned on the previous gesture, gesture style, and speech. It then reconstructs the gesture sequence via denormalization. Unlike generative models like GANs or Diffusion, VQ-VAE focuses more on compression rather than direct generation.
- **Diffusion**: Diffusion models focus on generating new data from noisy inputs by progressively denoising toward the original data. Diffusion-based approaches will be presented in Subsection 2.4.

2.3 Common Stages in Deep Learning Approaches for Gesture Generation

As presented in Subsection 1.1, gestures consist of sequences of 3D point coordinates. For each dataset, the number of bones per frame may vary.

Deep learning approaches to gesture generation are implemented using various techniques. However, we generalize the process into the following main stages, illustrated in Figure 5:

- (1) **Preprocessing**: In the preprocessing stage, data such as speech segments, gesture sequences, and text are

read and digitized into vectors or matrices that represent raw data information. Depending on the specific learning method, the selected initial data features may vary.

- (2) **Feature Processing**: In this stage, raw data such as speech and text are embedded into feature vectors. Different methods use different embedding models. The way gesture sequences are represented as feature vectors also varies across methods.
- (3) **Feature Extraction**: This stage uses linear transformation layers or CNN layers to extract features from the data. Text and speech features, after being processed, may be further passed through feature extraction layers to generate representation vectors corresponding to the input modalities.
- (4) **Feature Encoding**: In this stage, gesture, emotion, and speech vectors are encoded into a lower-dimensional latent space to facilitate learning the correlations among modalities in the feature fusion stage.
- (5) **Feature Fusion**: In this stage, features from speech, text, gestures, and other information are combined, typically using concatenation, fully connected layers, or operations such as vector addition or subtraction in the latent space.
- (6) **Feature Decoding**: In this stage, latent vectors are decoded or upsampled back to their original dimensionality.
- (7) **Rendering**: Once the output vectors are restored to their original size, they are converted back into BVH files and rendered using software such as Blender or Unity to visualize character motion.

2.4 Diffusion-based Model for Gesture Generation

- *MotionDiffuse* [Zhang et al. 2022] employs a conditional Diffusion model, with conditions based solely on text, excluding audio. Additionally, the model predicts noise rather than directly predicting the original gesture sequence. MotionDiffuse utilizes Self-Attention and Cross-Attention layers to model the correlation between textual features and gesture features during *Stage 5. Feature Fusion* (Figure 5).
- *Flame* [Kim et al. 2023] applies a Diffusion model with a Transformer-based architecture. In *Stage 2. Feature Processing* (Figure 5), it uses the pre-trained RoBERTa model to embed the text into textual feature vectors, which serve as the conditioning input. During *Stage 5. Feature Fusion* (Figure 5), the text is used as the CLS token prepended to the gesture sequence before passing through the Transformer Decoder. Similar to other methods, the model predicts the added noise rather than the original gesture sequence.

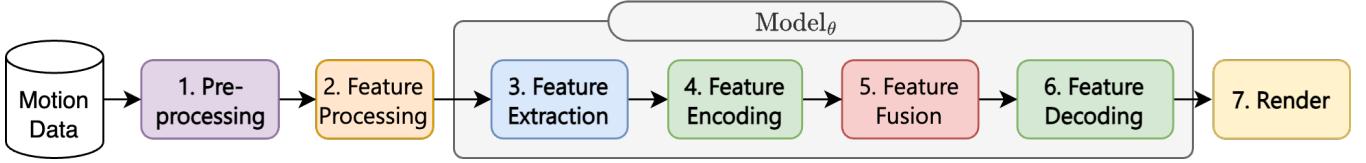


Figure 5: Common stages in gesture generation models.

- *DiffWave* [Kong et al. 2020] is a noise-predicting Diffusion model in which the time steps pass through multiple Fully Connected layers and a Swish activation function before feature fusion. It uses a dilated convolutional architecture inherited from WaveNet. DiffWave enables better representation of speech, improving the effectiveness of conditioning for the Diffusion model.
- *Listen, Denoise, Action* [Alexanderson et al. 2022] builds upon DiffWave [Kong et al. 2020], replacing the dilated convolution layers with a Transformer, and integrating Conformer modules to enhance model performance.
- *DiffSHEG* [Chen et al. 2024] employs a Diffusion model; in *Stage 2. Feature Processing*, it uses HuBERT to encode the audio signal. The model treats facial expressions as a signal for gesture generation and achieves real-time fusion of both facial expressions and gestures.
- *GestureDiffuCLIP* [Ao et al. 2023] uses a Diffusion model conditioned on text, leveraging Contrastive Learning with CLIP to integrate text features and control gesture styles. Similar to other prompt-based approaches such as StableDiffusion or Midjourney, it treats text as prompts to learn gestures from descriptive sentences.
- *Freetalker* [Yang et al. 2024] trains a Diffusion model on multiple datasets to generate speaker-specific gestures conditioned on speech and text. Unlike Transformer-based methods, Freetalker employs an Attention-based Network to model the correlation between textual, auditory, and gesture features during *Stage 5. Feature Fusion* (Figure 5).

2.4.1 Selected Diffusion-based Methods.

- *MDM* [Tevet et al. 2022] applies conditional Diffusion to gesture generation, using CLIP (Contrastive Language–Image Pre-training) embeddings of descriptive text as conditions. MDM adopts a Transformer-based architecture to reconstruct original gesture data. Like other text-based Diffusion approaches, in *Stage 3. Feature Extraction* (Figure 5), the input text is randomly masked to hide certain segments, enabling the model to learn the importance of each part for various gestures.
- In *Stage 5. Feature Fusion* (Figure 5), the text is prepended as a [CLS] token to the gesture sequence before passing through the Transformer Encoder, where self-attention

models the relationship between the text and each gesture frame. MDM predicts the original gesture data rather than noise.

- *DiffuseStyleGesture* [Yang et al. 2023a] extends MDM [Tevet et al. 2022] by incorporating audio, initial gestures, and style as conditioning inputs. In *Stage 1. Preprocessing* (Figure 5), the model processes coordinate vectors to obtain feature vectors of dimension $D = 1141$ per frame. In *Stage 2. Feature Processing* (Figure 5), DiffuseStyleGesture uses WavLM for audio embedding. In *Stage 5. Feature Fusion* (Figure 5), it improves upon MDM by applying Cross-Local Attention prior to the Transformer Encoder.

3 PROPOSED METHOD

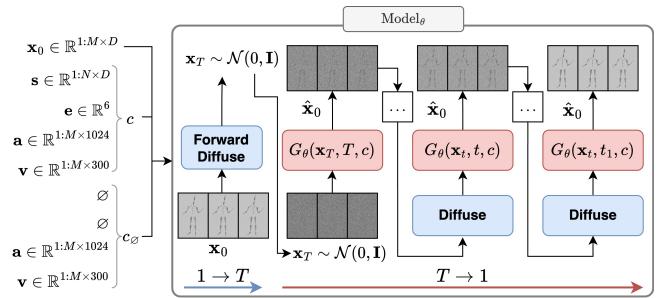


Figure 6: Overview of the DeepGesture model

The proposed **DeepGesture** model is based on the **DiffuseStyleGesture** model [Yang et al. 2023a], which applies the Diffusion model [Ho et al. 2020] with conditional guidance [Ho and Salimans 2022] (Classifier-Free Diffusion Guidance) to control features during the denoising process.

The similarities and differences of applying the diffusion model to the gesture generation task compared to image generation are as follows:

Similarities

- Uses the Diffusion model on gesture data $\mathbf{x}^{1:M \times D}$, with M temporal frames and $D = 1141$ representing motion coordinates per frame (analogous to image width and height).
- Uses conditional Diffusion with \mathbf{x}_0 objective.

- In stages 4. *Feature Encoding* and 6. *Feature Decoding* in Figure 5, the model uses a latent vector of dimension 256.

Differences

- Conditional gesture generation:
 - Emotional condition: $c = [s, e, a, v]$ and $c_\emptyset = [\emptyset, \emptyset, a, v]$.
 - Emotion state interpolation between e_1, e_2 using: $c = [s, e_1, a, v]$ and $c_\emptyset = [s, e_2, a, v]$.
- In stage 5. *Feature Fusion* Figure 5, the model uses Self-Attention: learning the relationship between emotions, seed gestures, and each frame (similar to DALL-E 2’s text-image alignment).
- In stage 5. *Feature Fusion* Figure 5, the model concatenates speech and text (analogous to ControlNet’s pixel-wise condition).

Here, \mathbf{x}_0 is a sequence of M gesture frames $\mathbf{x} \in \mathbb{R}^{1:M \times D}$ ($D = 1141$), with condition $c = [s, e, a, v]$ including seed gesture s , emotion e , speech a corresponding to the gesture, and text v .

The model’s objective is to learn parameters θ of the generative function G_θ with inputs being the noisy gesture matrix $\mathbf{x}_t \in \mathbb{R}^{1:M \times D}$, timestep t , and condition c . An overview of the proposed **DeepGesture** model is illustrated in Figure 6. As with standard diffusion models, it includes two processes: the diffusion process q and the denoising process p_θ with weights θ . The 1. *Preprocessing* stage will be presented in Subsection 4.2.

3.1 Feature Processing Stage

In the *Stage 2: Feature Processing* step (Figure 5), the goal is to convert raw input data into matrices or vectors suitable for input into the model.

- **Text $v \in \mathbb{R}^{1:M \times 300}$:** As discussed in Subsection 2.4, methods such as *MDM* [Tevet et al. 2022] and *DiffuseStyleGesture+* [Yang et al. 2023b] use gesture-descriptive prompts, similar to those used in Midjourney, as input for the model. However, such prompts are mainly used to cluster gestures rather than align them semantically. In contrast, we treat text as a semantic feature that aligns specific segments of text with corresponding gesture segments for digital human generation. Therefore, the contribution in this stage is to use speech data that has already been preprocessed (see Subsection 4.2) to obtain transcribed text. We use the FastText model [Bojanowski et al. 2017] to embed this text into vectors, which are then temporally aligned with the number of gesture frames, resulting in a text matrix $v \in \mathbb{R}^{1:M \times 300}$. For segments without text, zero vectors are used; for segments with vocabulary, the corresponding embedding is assigned for each gesture frame.

- **Speech $a \in \mathbb{R}^{1:M \times 1024}$:** All speech data in ‘.wav’ format is downsampled to 16 kHz. The speech corresponding to the gesture segment (4 seconds) is extracted as a waveform vector $a \in \mathbb{R}^{64000}$. Following DiffuseStyleGesture, this work uses the pre-trained WavLM Large model [Chen et al. 2022] to embed the raw waveform into a high-dimensional vector representing acoustic features. Linear interpolation is then applied to temporally align the latent features from WavLM at 20 fps, resulting in the speech matrix $a \in \mathbb{R}^{1:M \times 1024}$.
- **Emotion $e \in \mathbb{R}^6$:** Emotion is represented by one of six classes: Happy, Sad, Neutral, Old, Relaxed, and Angry. Each label is encoded using one-hot encoding to form the vector $e \in \mathbb{R}^6$.
- **Seed Gesture $s \in \mathbb{R}^{1:N \times D}$:** This is the initial gesture sequence composed of $N = 8$ frames, with each frame containing joint data for 75 joints. These are processed according to the formula in Equation 8 to yield a $D = 1141$ -dimensional vector.
- **Ground Truth Gesture $\mathbf{x}_0 \in \mathbb{R}^{1:M \times D}$:** This is the ground-truth gesture sequence of $M = 80$ frames (corresponding to 4 seconds at 20 fps), which is preprocessed to form the matrix $\mathbf{x}_0 \in \mathbb{R}^{1:M \times D}$.

3.2 Feature Extraction Stage

In the *Stage 3: Feature Extraction* step (Figure 5), the goal is to convert the input matrices into latent vectors that capture the semantic content of each modality. This is done by passing the feature data through linear transformation layers or Multilayer Perceptrons (MLPs).

- **Timestep $T \in \mathbb{R}^{256}$:** The timestep at each process step $t \in [0, T]$ enables the model to generalize the denoising process across different time steps. The objective is for the model to learn how the values should change with respect to t in order to accurately predict \mathbf{x}_0 . The timestep t is initialized using sinusoidal positional encoding: $\text{PE}(t) = [\sin\left(\frac{t}{10000^{2i/d}}\right), \cos\left(\frac{t}{10000^{2i/d}}\right)]$, and then passed through a Multilayer Perceptron (MLP) to obtain the vector $T \in \mathbb{R}^{256}$.
- **Speech $A \in \mathbb{R}^{1:M \times 64}$:** The speech feature matrix $a \in \mathbb{R}^{1:M \times 1024}$ is passed through a linear layer to reduce its dimensionality to a 64-dimensional feature vector, resulting in the matrix $A \in \mathbb{R}^{1:M \times 64}$.
- **Text $V \in \mathbb{R}^{1:M \times 64}$:** After preprocessing described in Subsection 4.2, the text is aligned to match the number of frames M , resulting in the matrix $v \in \mathbb{R}^{1:M \times 300}$. It is then passed through a linear transformation to reduce feature dimensionality, yielding the final matrix $V \in \mathbb{R}^{1:M \times 64}$, aligned with the speech matrix.

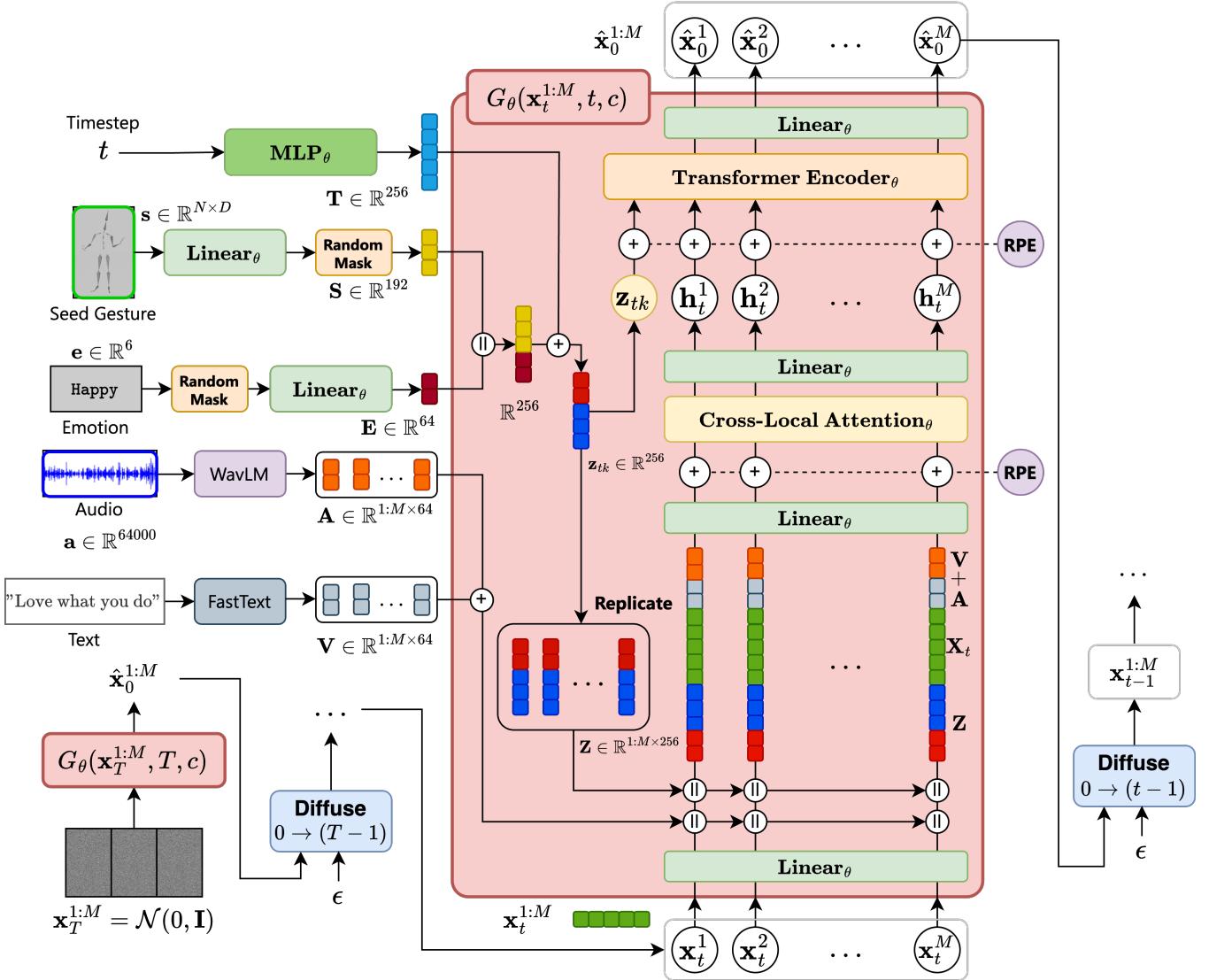


Figure 7: The Model Architecture in DeepGesture

- **Seed Gesture** $S \in \mathbb{R}^{192}$: The seed gesture $s \in \mathbb{R}^{1:N \times D}$ is processed through a linear transformation layer to obtain the vector $S \in \mathbb{R}^{192}$. During training, S is also passed through a random masking layer to randomly mask segments of gesture frames in N frames. This allows the model to learn how missing frames impact the final predicted gestures sequence.
- **Emotion** $E \in \mathbb{R}^{64}$: The emotion vector $e \in \mathbb{R}^6$ is passed through a linear transformation layer to produce the feature vector $E \in \mathbb{R}^{64}$. This is designed to be concatenated with the seed gesture vector $S \in \mathbb{R}^{192}$ to form a 256-dimensional latent vector.

- **Noisy Gesture** $x_T \in \mathbb{R}^{1:M \times D}$: During training, x_t is the noisy gesture that shares the same dimensionality as the original gesture x_0 and is sampled from a standard normal distribution $\mathcal{N}(0, I)$. The initial noisy gesture x_T is drawn from a Gaussian distribution, and the subsequent x_t , $t < T$ are produced through iterative noising steps, as illustrated in Figure 6.

3.3 Feature Encoding Stage

In the *Stage 4: Feature Encoding* step (Figure 5), the objective is to reduce the dimensionality of the input data to a lower latent space, in order to alleviate computational overhead and avoid the explosion of processing complexity.

The primary data used in the diffusion process is the gesture sequence $\mathbf{x}_t \in \mathbb{R}^{1:M \times D}$. As illustrated in Figure 7, the gesture sequence with size $M \times D$ is passed through a linear transformation layer Linear_θ to produce the matrix $\mathbf{X} \in \mathbb{R}^{1:M \times 256}$. This dimensionality reduction of \mathbf{x} is performed prior to passing the gesture sequence through the Cross-Local Attention and Transformer Encoder layers, in order to compute correlations across multiple modalities.

3.4 Feature Fusion Stage

In the *Stage 5: Feature Fusion* step (Figure 5), the goal is to compute inter-feature correlations using concatenation, addition, or attention-based mechanisms.

First, the seed gesture vector $\mathbf{S} \in \mathbb{R}^{192}$ and the emotion vector $\mathbf{E} \in \mathbb{R}^{64}$ are concatenated to form a single vector of size 256, since 256 is the chosen hidden dimensionality for computing feature correlations. This vector is then added to the timestep vector \mathbf{T} to form the final vector $\mathbf{z}_{tk} \in \mathbb{R}^{256}$.

$$\mathbf{z}_{tk} = \text{concat}(\mathbf{E} \parallel \mathbf{S}) + \mathbf{T} \quad (1)$$

The feature fusion process of \mathbf{z}_{tk} is illustrated in Figure 9a.

3.4.1 Frame-wise Feature Integration. Next, $\mathbf{z}_{tk} \xrightarrow{\text{replicate}} \mathbf{Z}$ is replicated M times to match the dimensionality of M frames, resulting in the matrix $\mathbf{Z} \in \mathbb{R}^{1:M \times 256}$, as shown in Figure 7.

The speech feature matrix \mathbf{A} and the text feature matrix \mathbf{V} are then added together to produce a matrix that combines both modalities. This combined matrix is then concatenated with the gesture feature matrix \mathbf{X} . Finally, the result is concatenated with matrix \mathbf{Z} to obtain the full feature matrix \mathbf{M} .

$$\mathbf{M} = \text{concat}(\mathbf{Z} \parallel \text{concat}(\mathbf{X} \parallel (\mathbf{V} + \mathbf{A}))) \quad (2)$$

The matrix $\mathbf{M} \in \mathbb{R}^{1:M \times P}$, as shown in Equation 2, represents the frame-wise feature matrix from frame 1 to frame M , where each frame has dimensionality P , which is the sum of all concatenated feature vectors. Given $\mathbf{X} \in \mathbb{R}^{1:M \times 256}$, $\mathbf{Z} \in \mathbb{R}^{1:M \times 256}$, and $\mathbf{A}, \mathbf{V} \in \mathbb{R}^{1:M \times 64}$, the resulting dimensionality is $P = 256 + 256 + 64$.

Subsequently, the matrix \mathbf{m} is passed through a linear transformation to reduce its dimensionality from P to 256, resulting in the matrix $\mathbf{m}_t \in \mathbb{R}^{1:M \times 256}$.

$$\mathbf{m}_t = \text{Linear}_\theta(\mathbf{M}) \quad (3)$$

3.4.2 Attention Mechanism in the Feature Integration Process. In the proposed model, the attention mechanism [Vaswani et al. 2017] is employed to integrate features. The objective of applying attention is to capture the correlation between individual frames in the sequence. The attention mechanism is utilized in both the Cross-Local Attention and the Self-Attention layers within the Transformer Encoder.

The attention mechanism is formulated as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \text{Mask}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T + \text{Mask}}{\sqrt{C}}\right)\mathbf{V} \quad (4)$$

In the Attention formula above, \mathbf{Q} (Query), \mathbf{K} (Key), and \mathbf{V} (Value) are matrices obtained by passing the input through linear transformation matrices: $\mathbf{Q} = \mathbf{X}\mathbf{W}_Q$, $\mathbf{K} = \mathbf{X}\mathbf{W}_K$, $\mathbf{V} = \mathbf{X}\mathbf{W}_V$. The input \mathbf{X} is a matrix representing a sequence of M frames, where each frame is a concatenated vector composed of various feature vectors, including seed gestures, text, speech, emotion, and the gesture \mathbf{x}_t that we aim to denoise. The term \sqrt{C} is a normalization constant that accounts for the dimensionality of the matrices.

The Local-Cross Attention mechanism is controlled to focus only on local motion features of gestures and features in neighboring frames.

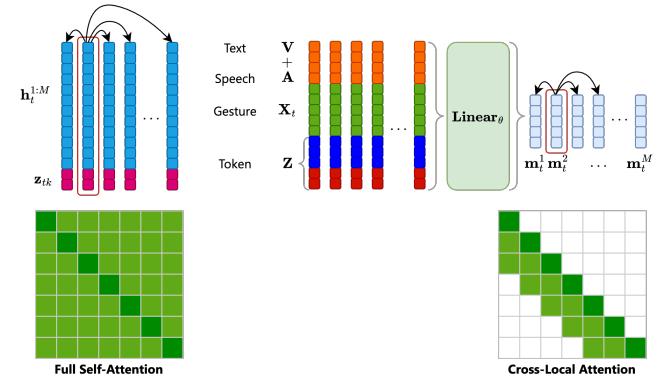


Figure 8: Attention Mechanism in Transformer Encoder and Cross-Local Attention

The Attention mechanism functions like a dictionary, where the final retrieved information is the matrix \mathbf{V} (value), while \mathbf{Q} (query) represents the keyword being searched for, and \mathbf{K} (key) is the set of keywords in the lookup dictionary. The Attention process computes the similarity between \mathbf{Q} and \mathbf{K} to determine the weights for the values in \mathbf{V} .

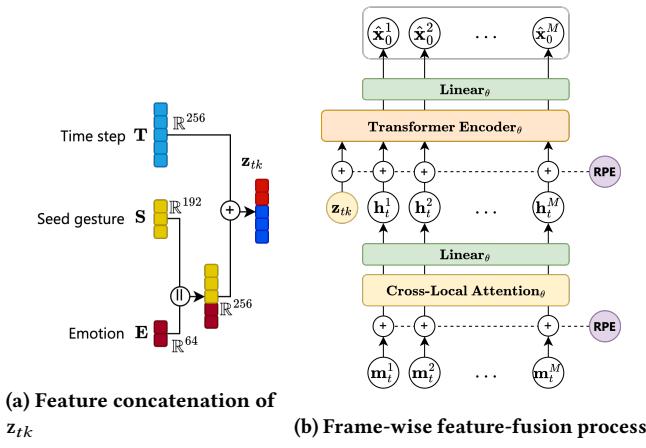
The final result is a weighted combination of the values in \mathbf{V} , where the values corresponding to keys most similar to the query receive higher weights. \mathbf{M} is the mask used to perform local attention. The Cross-Local Attention mechanism is illustrated on the right in Figure 8, while the Transformer Encoder layer uses Self-Attention shown on the left.

3.4.3 Combining Local Features with Cross-Local Attention. The matrix $\mathbf{m}_t \in \mathbb{R}^{1:M \times 256}$ is then passed through the Cross-Local Attention layer to compute the correlations between local features.

$$\mathbf{h}_t = \text{Linear}_\theta(\text{Cross-Local Attention}(\mathbf{m}_t)) \quad (5)$$

Cross-local Attention is performed with $Q = K = V = \mathbf{m}_t$. Following the idea of the Routing Transformer method [Roy et al. 2021], Cross-local Attention highlights the importance of constructing intermediate feature-vector representations before passing them through the Transformer Encoder layer, as shown in Figure 7. The feature vectors are augmented with a relative position-encoding vector **RPE** (Relative Position Encoding) to preserve temporal ordering before entering the Cross-Local Attention layer.

After Cross-Local Attention, the model is forwarded through a linear layer, as in Equation 5, to align with the M frames and obtain the matrix $\mathbf{h}_t \in \mathbb{R}^{1:M \times D}$.



3.4.4 Global Feature Fusion with the Transformer Encoder. Following MDM [Tevet et al. 2022], the vector \mathbf{z}_{tk} is the first token that encodes information for the entire frame sequence, analogous to the CLS token in BERT [Devlin et al. 2019], which summarizes an entire text segment. Here, we use \mathbf{z}_{tk} , with $\mathbf{z}_{tk} \in \mathbb{R}^{256}$ (Equation 1), as the first token representing global features for the whole sequence of M frames.

$$\mathbf{X}_0 = \text{Transformer Encoder}(\text{concat}(\mathbf{z}_{tk} \parallel \mathbf{h}_t^{1:M})) \quad (6)$$

The vectors \mathbf{h}_t represent the sequence of M frames. Similar to Reformer [Kitaev et al. 2020], before entering the Self-Attention layer of the Transformer Encoder, the model employs Relative Position Encoding (RPE) instead of absolute position encoding, improving efficiency on long sequences. Within the Transformer Encoder layer [Vaswani et al. 2017], relationships among the data sequences are computed. The Transformer Encoder applies the same self-attention mechanism as in Equation 4 but without the Mask, enabling correlations across the entire sequence to be captured.

3.5 Feature Decoding Stage

In stage 6. *Feature Decoding* (Figure 5), once feature correlations are computed, the goal is to upsample the data back to its original dimensionality.

As illustrated in Figure 7, the latent matrix \mathbf{X}_0 , after passing through the Transformer Encoder to capture correlations among heterogeneous data types, is fed into a linear projection layer $\hat{\mathbf{x}}_0 = \text{Linear}_\theta(\mathbf{X}_0)$ to restore the latent matrix to its original size, yielding $\hat{\mathbf{x}}_0 \in \mathbb{R}^{1:M \times D}$.

The final rendering step is presented in Subsection 4.4.

3.6 Emotion Control in Gesture Generation

The preceding steps enable the model to learn gesture generation. To incorporate emotions across different contexts, each emotion is parameterized and varied so that the predictions faithfully express the designated affect.

Analogous to conditional denoising models [Ho and Salimans 2022; Tevet et al. 2022], we use the condition vector $c = [\mathbf{s}, \mathbf{e}, \mathbf{a}, \mathbf{v}]$, where \mathbf{s} is the seed gesture, \mathbf{e} the emotion, \mathbf{a} the associated speech, and \mathbf{v} the text. The conditional diffusion model injects c at every timestep t in the denoising network $G_\theta(\mathbf{x}_t, t, c)$, with $c_\emptyset = [\emptyset, \emptyset, \mathbf{a}, \mathbf{v}]$ (unconditional) and $c = [\mathbf{s}, \mathbf{e}, \mathbf{a}, \mathbf{v}]$ (conditional). A random mask applied to the seed-gesture and emotion vectors conveniently switches labels, allowing optimization under diverse conditions.

$$\hat{\mathbf{x}}_{0,c,c_\emptyset,y} = \gamma G(\mathbf{x}_t, t, c) + (1 - \gamma) G(\mathbf{x}_t, t, c_\emptyset) \quad (7)$$

Classifier-free guidance [Ho and Salimans 2022] further enables interpolation between two emotions \mathbf{e}_1 and \mathbf{e}_2 by setting $c = [\mathbf{s}, \mathbf{e}_1, \mathbf{a}, \mathbf{v}]$ and $c_\emptyset = [\mathbf{s}, \mathbf{e}_2, \mathbf{a}, \mathbf{v}]$:

$$\hat{\mathbf{x}}_{0,\gamma,c_1,c_2} = \gamma G(\mathbf{x}_t, t, c_1) + (1 - \gamma) G(\mathbf{x}_t, t, c_2).$$

3.7 Training Procedure

Algorithm 1 Training in DeepGesture

- (1) Pre-compute $\gamma, \sqrt{\alpha_t}, \sqrt{1 - \alpha_t}, \sqrt{\bar{\alpha}_t}$, and random noise ϵ_t for each timestep $t : 1 \rightarrow T$. Define the noise schedule $\{\alpha_t \in (0, 1)\}_{t=1}^T$.
- (2) Sample the initial label \mathbf{x}_0 from the normalized data distribution.
- (3) Randomly generate Bernoulli masks $c_1 = [\mathbf{s}, \mathbf{e}_1, \mathbf{a}, \mathbf{v}]$, $c_2 = [\mathbf{s}, \mathbf{e}_2, \mathbf{a}, \mathbf{v}]$, or $c_2 = [\emptyset, \emptyset, \mathbf{a}, \mathbf{v}]$.
- (4) Add noise to obtain the noisy gesture \mathbf{x}_t :

$$\mathbf{x}_t = \sqrt{\alpha_t} \mathbf{x}_0 + \sqrt{1 - \alpha_t} \epsilon_t.$$

- (5) Sample t uniformly from $[1, T]$.
- (6) Given \mathbf{x}_t , t , and masks c_1, c_2 , predict the gesture sequence:
- $\hat{\mathbf{x}}_{0,\gamma,c_1,c_2} = \gamma G_\theta(\mathbf{x}_t, t, c_1) + (1 - \gamma) G_\theta(\mathbf{x}_t, t, c_2)$.
- (7) Compute the loss and gradient to update θ :
- $\mathcal{L}^t = \mathbb{E}_{t, \mathbf{x}_0, \epsilon_t} [\text{HuberLoss}(\mathbf{x}_0, \hat{\mathbf{x}}_0)]$.
- (8) Repeat from step 6 until convergence, obtaining the optimal parameters θ' .

Algorithm 1 trains the DeepGesture model by first computing the required values and hyper-parameters $\gamma, \sqrt{\alpha_t}$,

$\sqrt{1 - \alpha_t}$, $\sqrt{\bar{\alpha}_t}$, and ϵ_t – for every timestep t ($1 \dots T$). The initial label \mathbf{x}_0 , representing the ground-truth gesture, is drawn from the normalized data distribution. Random Bernoulli masks c_1 and c_2 emulate different conditions (gesture, emotion, speech, or text), with one mask possibly lacking emotion information. Noise is then added to create the noisy gesture \mathbf{x}_t . A timestep t is sampled uniformly, and \mathbf{x}_t with the masks is fed into the model to predict the original gesture sequence as a weighted combination of conditional outputs. The Huber loss between ground-truth and prediction is used to update θ . This cycle repeats until the model converges, yielding the optimal parameters θ' .

3.8 Sampling Process

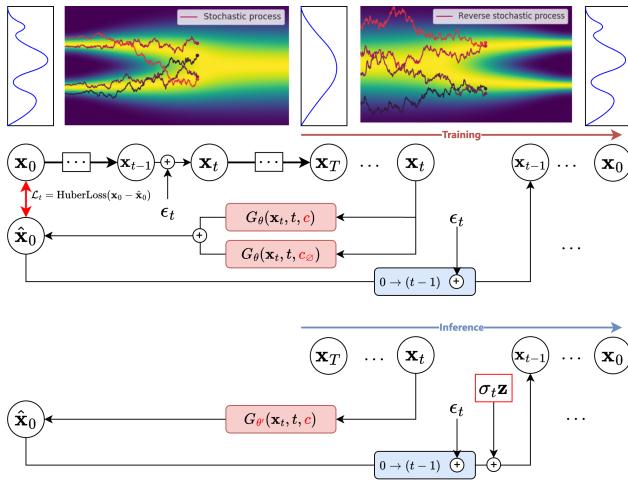


Figure 10: Offline (Training) and Online (Inference) Phases

To generate gestures of arbitrary length, the original sequence is segmented into clips of length M . During training, the seed gesture can be chosen by randomly selecting a gesture from the dataset or by averaging the clipped segments—here, the mean rotation angles are used. Generated frames are processed sequentially, with the last $N = 8$ frames taken as the seed for the next iteration. For each clip, the gesture \mathbf{x}_t is denoised via $\hat{\mathbf{x}}_0 = G_{\theta'}(\mathbf{x}_t, t, c)$; noise is re-added to obtain \mathbf{x}_{t-1} , and the procedure repeats until $t = 1$, yielding \mathbf{x}_0 .

Algorithm 2 starts by initializing the noisy gesture \mathbf{x}_T from $\mathcal{N}(0, \mathbf{I})$. The values $\sqrt{\alpha_t}$, $\sqrt{1 - \alpha_t}$, and $\sqrt{\bar{\alpha}_t}$ obtained during training, together with σ_t , are employed at each timestep ($1 \dots T$). Each 4-second speech segment is represented by \mathbf{a} , and the seed gesture \mathbf{s} is taken as the data mean or from the previously inferred segment. The desired emotion and the transcript form the condition $c = [\mathbf{s}, \mathbf{e}, \mathbf{a}, \mathbf{v}]$. The algorithm proceeds sequentially from T to 1 : random noise \mathbf{z} is generated, the model predicts $\hat{\mathbf{x}}_0^{(t)}$ from \mathbf{x}_t , t , and c , then $\hat{\mathbf{x}}_{t-1}^{(t)}$ is

Algorithm 2 Sampling in DeepGesture

- (1) Initialize with noise: $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$.
 - (2) Retrieve $\sqrt{\alpha_t}$, $\sqrt{1 - \alpha_t}$, and $\sqrt{\bar{\alpha}_t}$ from training; precompute σ_t from α_t for each timestep $t : 1 \rightarrow T$.
 - (3) Split each 4-second speech segment into $\mathbf{a} \in \mathbb{R}^{64000}$. The initial seed gesture \mathbf{s} is the data mean and later updated from the inferred gesture segment. Select the desired emotion, obtain the transcript \mathbf{v} from speech \mathbf{a} , and form the condition $c = [\mathbf{s}, \mathbf{e}, \mathbf{a}, \mathbf{v}]$.
 - (4) For each timestep, take t **sequentially** from $[T, \dots, 1]$.
 - (5) Sample random noise $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$.
 - (6) Infer $\hat{\mathbf{x}}_0^{(t)} = G_{\theta'}(\mathbf{x}_t, t, c)$.
 - (7) Diffuse $\hat{\mathbf{x}}_0^{(t)}$ from step 0 $\rightarrow t$ to obtain $\hat{\mathbf{x}}_{t-1}^{(t)}$.
 - (8) Add noise: $\hat{\mathbf{x}}_{t-1} = \hat{\mathbf{x}}_{t-1}^{(t)} + \sigma_t \mathbf{z}$.
 - (9) Return to step 4. When $t = 1$, output the denoised gesture $\hat{\mathbf{x}}_0$.
-

computed and perturbed with noise. This loop continues until $t = 1$, after which the algorithm outputs the final denoised gesture $\hat{\mathbf{x}}_0$.

4 EXPERIMENTS AND RENDERING

4.1 Dataset

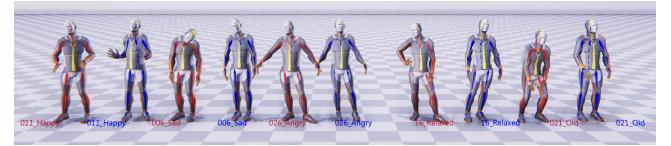


Figure 11: Gesture samples of difference emotion

We use the ZeroEGGS retarget dataset [Ghorbani et al. 2023], [Ghorbani et al. 2022], a motion capture dataset designed for research and development of gesture generation models. It includes 67 monologue clips performed by a female motion capture actor, with a total duration of 135 minutes. The monologues are annotated with 6 different emotions: Happy, Sad, Neutral, Old, Relaxed, and Angry, enabling simulation of various emotional states in gestures and body movements. ZeroEGGS provides a rich platform for studying the integration of speech and dynamic gestures, supporting the development of models capable of adapting gestures according to emotions and text semantics.

4.2 Data Preprocessing

In stage 1. Data Preprocessing (Figure 5), gesture, speech, and text data are read and processed to be represented as vectors or matrices containing information derived from raw data.

For **text data**: we use the nltk library for tokenization and contractions to normalize contracted words.

One of our contributions is converting the speech data available in ZeroEGGS using Adobe Speech To Text, then align the phonetic timestamps using the Montreal Forced Aligner [Saxon et al. 2020] with an English phoneme dictionary to match the gesture frame rate, generating TextGrid files. From these TextGrids containing word-level timing information, we use gensim to generate word2vec embeddings.

Gesture data consists of BVH (BioVision Motion Capture) files captured via motion capture systems. BVH files include two components: Hierarchy and Motion. Specifically:

- **HIERARCHY:** defines a skeletal tree containing 75 bones $\{b_1, b_2 \dots b_{75}\}$, each with an initial position (OFFSET) and CHANNELS parameters specifying the type and order of rotation angles (Zrotation, Yrotation, Xrotation) and position (Xposition, Yposition, Zposition), which are defined in the MOTION section. The first bone (usually Hips) is the root bone b_{root} , used to define the T-pose via forward kinematics as the initial pose of the skeleton before applying motion.
- **MOTION:** a sequence of frames, each containing motion data representing changes of all 75 bones as defined by the CHANNELS in the HIERARCHY.

Our model converts Euler rotation angles to quaternion rotation angles, where a quaternion is a 4-dimensional vector.

$$\mathbf{g} = \left[p_{root}, r_{root}, p'_{root}, r'_{root}, p_{joins}, r_{joins}, p'_{joins}, r'_{joins}, d_{gaze} \right] \quad (8)$$

Here, each $\mathbf{g} \in \mathbb{R}^{1141}$ includes:

- $p_{root} \in \mathbb{R}^3$: coordinates of the root joint
- $r_{root} \in \mathbb{R}^4$: rotation (quaternion) of the root joint
- $p'_{root} \in \mathbb{R}^3$: velocity of the root position
- $r'_{root} \in \mathbb{R}^3$: angular velocity of the root rotation
- $p_{joins} \in \mathbb{R}^{3n_{join}}$: positions of other joints
- $r_{joins} \in \mathbb{R}^{6n_{join}}$: joint rotations on the X and Y planes
- $p'_{joins} \in \mathbb{R}^{3n_{join}}$: velocity of joint positions
- $r'_{joins} \in \mathbb{R}^{3n_{join}}$: angular velocity of joint rotations
- $d_{gaze} \in \mathbb{R}^3$: gaze direction

The original gesture sequences in Euler angles are converted to radians, then converted from Euler to Quaternion as detailed in Subsection A.3.

Speech data: $a_{raw} \in \mathbb{R}^{\text{length}}$ is raw speech sampled at 16000 Hz, trimmed into segments $a \in \mathbb{R}^{64000}$ corresponding to 4 seconds. The paper uses ffmpeg-normalize to normalize volume to a level lower than the original.

Emotion: Emotion data is represented using a predefined one-hot encoded vector. During sampling, the filename encodes the target emotion.

All data is stored using the h5 format.

4.3 Training Process

The entire model training process was conducted over approximately two weeks with the following parameters: number of training steps $T = 1000$, using an Nvidia 3090 GPU. The learning rate was set to 3×10^{-5} , batch size was 640, and a total of 43,853 samples were trained. At each step, t is randomly sampled and input to f_θ to predict \mathbf{x}_0 . The emotional control parameter was set to $\gamma = 0.1$. The probability of applying random masking to the emotion and initial gesture matrices was 10%, using a Bernoulli distribution to randomly hide/reveal these matrices.

The β parameter was scheduled linearly from 0.5 → 0.999. The HuberLoss($\mathbf{x}_0, \hat{\mathbf{x}}_0$) is computed as follows:

- If $|\mathbf{x}_0 - \hat{\mathbf{x}}_0| \leq \delta$ then $\mathcal{L}_{\delta, \mathbf{x}_0, \hat{\mathbf{x}}_0} = \frac{1}{2}(\mathbf{x}_0 - \hat{\mathbf{x}}_0)^2$: Below the threshold, the loss is computed as squared distance (similar to MSE), which is sensitive to small errors and provides smooth gradients.
- If $|\mathbf{x}_0 - \hat{\mathbf{x}}_0| > \delta$ then $\mathcal{L}_{\delta, \mathbf{x}_0, \hat{\mathbf{x}}_0} = \delta \cdot |\mathbf{x}_0 - \hat{\mathbf{x}}_0| - \frac{1}{2}\delta^2$: Above the threshold, the loss behaves like MAE, reducing sensitivity to large errors and improving robustness against outliers.

The training process is implemented in the open-source repository: Github/DeepGesture¹.

4.4 Rendering Process in Unity

To visualize the gesture generation process from model output, we use Unity in stage 7. *Rendering* (Figure 5), extending code from the DeepPhase model [Starke et al. 2022]. The generated output is in BVH (BioVision Motion Capture) format. In Unity, the author adds C-Sharp scripts to render gestures based on coordinates and labels, with bone positions and rotations represented using quaternions.

Rendering details are presented in Appendix B.

The Unity project source code is available at: Github/DeepGesture-Unity².

5 EVALUATION AND RESULTS

5.1 Evaluation Methods

The evaluation is conducted using two primary metrics: Mean Opinion Scores (MOS) and Fréchet Gesture Distance (FGD).

5.1.1 Human Perception Evaluation. Mean Opinion Scores (MOS)

As of now, there is no standardized metric for gesture generation, particularly in the speech-driven context. Consequently, this work relies on subjective human evaluations, a practice consistent with prior research [Alexanderson et al.

¹<https://github.com/DeepGesture/DeepGesture>

²<https://github.com/DeepGesture/deepgesture-unity>

2022; Kucherenko et al. 2021; Yoon et al. 2022], since objective metrics still fail to fully capture human perception.

MOS is assessed based on the following three criteria:

- Human-likeness
- Gesture-Speech Appropriateness
- Gesture-Style Appropriateness

A notable contribution of this work is the development of the GENEVA Leaderboard [Nagy et al. 2024], which includes HEMVIP (**H**uman **E**valuation of **M**ultiple **V**ideos in **P**arallel). HEMVIP is designed to compare gesture generation quality between different models using rendered videos.

Within the GENEVA consortium (**G**eneration and **E**valuation of **N**on-verbal **B**ehaviour for **E**mbodied **A**gents), evaluators are recruited via Prolific. Participants watch videos rendered by different models and assign ratings of *Left Better*, *Equal*, or *Right Better*. These responses are converted into numerical scores of -1 , 0 , and 1 , respectively, and updated via the Elo rating system.

The evaluation platform and source code are available at github.com/hemvip/hemvip.github.io³.

5.1.2 Objective Metric Evaluation. Mean Square Error (MSE)

We compute the mean square error between the predicted gesture sequence $\hat{\mathbf{y}}_i^{1:M \times D}$ and the corresponding ground-truth sequence $\mathbf{y}_i^{1:M \times D}$ as:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n \|\mathbf{y}_i^{1:M \times D} - \hat{\mathbf{y}}_i^{1:M \times D}\|^2 \quad (9)$$

Where:

- n is the number of data samples,
- M is the number of frames per sample,
- D is the number of dimensions per frame,
- $\|\cdot\|^2$ is the squared Frobenius norm.

A lower MSE indicates a smaller difference between predicted and ground-truth gestures. Evaluation results are provided in Subsubsection 5.2.2.

Fréchet Gesture Distance (FGD)

Inspired by the Fréchet Inception Distance (FID) used in image generation, the Fréchet Gesture Distance (FGD) measures the similarity in distribution between predicted gestures $\hat{\mathbf{y}}$ and real gestures \mathbf{y} :

$$\text{FGD} = \|\hat{\mu} - \mu\|^2 + \text{Tr}(\Sigma + \hat{\Sigma} - 2\sqrt{\Sigma\hat{\Sigma}}) \quad (10)$$

Where:

- μ and $\hat{\mu}$ are the empirical means of real and generated data,
- Σ and $\hat{\Sigma}$ are the corresponding covariance matrices,
- $\text{Tr}(\cdot)$ denotes the trace operator,
- $\sqrt{\Sigma\hat{\Sigma}}$ is the matrix square root of the product.

³HEMVIP 2: <https://github.com/hemvip/hemvip.github.io>

Lower FGD indicates higher distributional similarity, thus better quality. Evaluations are conducted on $\hat{\mathbf{x}}^0, \mathbf{x}_0 \in \mathbb{R}^{1:M \times D}$.

5.2 Evaluation Results

5.2.1 User Study Results. MOS Evaluation

Due to the high cost and resource requirements of human evaluations, this work reuses the subjective evaluation results from **DiffuseStyleGesture** [Yang et al. 2023a] as a baseline. The proposed **OHGesture** model is not included in the subjective study.

To assess visual quality, a separate user study is conducted comparing the proposed method with real motion capture data. Each evaluated clip is 11–51 seconds long (mean: 31.6s), exceeding the typical GENEVA clip duration of 8–10s [Yoon et al. 2022]. Longer clips offer clearer insight into gesture consistency [Yang et al. 2022]. Ratings range from 5 (excellent) to 1 (bad).

Model	Human Likeness ↑	Gesture-Speech Appropriateness ↑
Ground Truth	4.15 ± 0.11	4.25 ± 0.09
Ours	4.11 ± 0.08	4.11 ± 0.10

Table 1: Mean Opinion Score (MOS) evaluation results

5.2.2 Quantitative Results. MSE Evaluation

MSE is computed on gesture sequences across M frames and reported across six emotion categories.

Emotion	Neutral	Sad	Happy	Relaxed	Elderly	Angry
DiffuseStyleGesture	75.04	51.40	110.18	130.83	116.03	78.53
ZeroEGG	136.33	81.22	290.47	140.24	102.44	181.07
OHGesture	161.22	89.58	279.95	156.93	99.86	215.24

Table 2: MSE results across six emotion categories

FGD Evaluation

This work introduces FGD and implements the open-source tool GestureScore⁴. GestureScore encodes gesture frames via an Inception V3-based model into a 32×32 latent representation. FGD is then computed via Equation 10.

Model Variant	FGD on Feature Vectors	FGD on Raw Data
Ground Truth	-	-
OHGesture (Feature D=1141)	2.058	9465.546
OHGesture (Rotations)	3.513	9519.129

Table 3: FGD results for OHGesture on $\mathbf{x}^{1:M \times D}$

- **Feature Vectors:** Skeletons from BVH files are encoded into vectors with $D = 1141$ dimensions (Equation 8).
- **Rotations:** Joint rotations are extracted, resulting in $D = 225$ features (75 joints $\times 3$ rotation angles).

⁴<https://github.com/GestureScore/GestureScore>

6 STRENGTHS AND LIMITATIONS

The proposed DeepGesture model introduces several key advantages that advance the development of more natural and expressive human-machine interaction systems. Nevertheless, certain limitations remain and represent important directions for future work.

Strengths:

- **High realism:** DeepGesture generates gestures that exhibit strong human-likeness, closely mirroring the timing and rhythm of natural speech. The model effectively synchronizes gestures with both the emotional tone and semantic content of speech.
- **Robust generalization:** Leveraging the denoising capabilities of diffusion models, DeepGesture can generate gestures for speech and emotional contexts not encountered during training, indicating strong potential for real-world deployment across varied domains.
- **Multidimensional controllability:** The diffusion-based architecture supports flexible control over multiple attributes, including the interpolation and modulation of emotional states, enabling expressive and context-sensitive gesture synthesis.

Limitations:

- **Lack of real-time inference:** The current implementation is not optimized for real-time execution and requires multi-step generation followed by offline rendering.
- **Suboptimal motion representation:** The gesture feature representation with $D = 1141$ dimensions is processed as a 2D image structure, which may not fully capture temporal motion dynamics and biomechanical features.
- **Sensitivity to input quality:** The model's performance heavily depends on clean, high-quality speech input. In cases of noisy or emotionally ambiguous audio, the gesture output may be less accurate or less expressive.

7 CONCLUSION

7.1 Achieved Results

This work presents DeepGesture, a gesture generation model that achieves high realism and naturalness through diffusion-based modeling. A primary contribution is its precise synchronization between generated gestures and the emotional content of speech, extending generalization beyond the training distribution. The model demonstrates the ability to produce context-aware gestures even in low-probability speech scenarios.

Another major advancement lies in expanding input modalities. In addition to speech, gesture, and emotion labels, the

model incorporates textual features obtained through automatic transcription. This multimodal approach enables the system to capture the semantic context of input speech more effectively and generate more appropriate gestures.

7.2 Future Work

Several promising directions can further improve DeepGesture and expand its applicability in real-world settings:

- **Real-time inference optimization:** Future efforts will focus on transforming DeepGesture into a real-time gesture generation system by reducing dependencies on offline rendering tools (e.g., Unity) and optimizing latency for interactive applications.
- **Efficient sampling:** The current diffusion process involves many sampling steps. Reducing these without compromising generation quality remains an important area for improving system responsiveness.
- **Advanced embeddings:** Incorporating richer embedding techniques—potentially combining speech, text, prosody, and affective signals—could further enhance gesture appropriateness across different languages and cultural contexts.
- **Multilingual generalization:** Extending the model to support multiple languages will broaden its applicability and ensure culturally appropriate gesture behavior.
- **Integration with phase-aware models:** A future integration with the DeepPhase model [Starke et al. 2022] is planned to improve the representation of temporal dynamics and support real-time applications via phase-informed motion generation.
- **Improved automatic metrics:** To reduce reliance on subjective human evaluation, ongoing work aims to design robust automatic evaluation metrics that better reflect human perception and can serve as internal feedback mechanisms during training and inference.

7.3 Closing Remarks

Through experimental validation and qualitative analysis, the DeepGesture model – extending DiffuseStyleGesture—demonstrates the ability to generate realistic gestures for both in-distribution and out-of-distribution speech, including synthetic voices such as that of Steve Jobs (see Appendix B). This highlights the promise of diffusion-based approaches in modeling complex, expressive, and low-frequency gesture behaviors.

Additionally, we contribute open-source code, including rendering pipelines and data processing tools built on Unity, available on GitHub. These resources provide a solid foundation for future development and reproducibility. The integration of text alongside speech and emotion into the gesture

generation pipeline marks an important step toward building fully multimodal agents, capable of more intuitive and human-like interactions in diverse application domains.

Acknowledgments

This research is partially supported by OpenHuman AI. We thank Daniel Holden for providing the retargeting dataset [Ghorbani et al. 2023]. We also acknowledge the use of the Ubisoft La Forge ZeroEGGS Animation Dataset, which includes a version exported in FBX and BVH formats.

References

- Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-controllable speech-driven gesture synthesis using normalising flows. *Computer Graphics Forum* 39, 2 (2020), 487–496.
- Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. 2022. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *CoRR* abs/2211.09707 (2022).
- Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. 2022. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Trans. Graph.* 41, 6 (2022), 209:1–209:19.
- Tenglong Ao, Zeyi Zhang, and Libin Liu. 2023. Gesturediffuclip: Gesture diffusion model with clip latents. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–18.
- Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. 2021. Speech2affectvegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2027–2036.
- 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.
- Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*. 413–420.
- Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. 2024. Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7352–7361.
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (Oct. 2022), 1505–1518. doi:10.1109/jstsp.2022.3188113
- Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. 2015. Predicting co-verbal gestures: A deep and temporal modeling approach. In *Intelligent Virtual Agents: 15th International Conference, IVA 2015, Delft, The Netherlands, August 26–28, 2015, Proceedings 15*. Springer, 152–166.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805* [cs.CL] <https://arxiv.org/abs/1810.04805>
- Paul Ekman and Wallace V Friesen. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *semiotica* 1, 1 (1969), 49–98.
- Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F. Troje, and Marc-André Carboneau. 2022. ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech. *arXiv:2209.07556* [cs.GR] <https://arxiv.org/abs/2209.07556>
- Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F. Troje, and Marc-André Carboneau. 2023. ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech. *Computer Graphics Forum* 42, 1 (2023), 206–216. [arXiv:https://onlinelibrary.wiley.com/doi/10.1111/cgf.14734](https://onlinelibrary.wiley.com/doi/10.1111/cgf.14734) doi:10.1111/cgf.14734
- Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. 2020. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–14.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
- Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).
- Chien-Ming Huang and Bilge Mutlu. 2012. Robot behavior toolkit: generating effective social behaviors for robots. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. 25–32.
- Jihoon Kim, Jiseob Kim, and Sungjoon Choi. 2023. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 8255–8263.
- Michael Kipp. 2005. *Gesture generation by imitation: From human behavior to computer character animation*. Universal-Publishers.
- Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451* (2020).
- Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2020. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761* (2020).
- 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.
- 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 GENEVA Challenge 2020. In *26th international conference on intelligent user interfaces*. 11–21.
- Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. 2010. Gesture controllers. In *Acm siggraph 2010 papers*. ACM, 1–11.
- Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022b. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European conference on computer vision*. Springer, 612–630.
- Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. 2022a. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10462–10472.
- Evelyn McClave. 1994. Gestural beats: The rhythm hypothesis. *Journal of psycholinguistic research* 23, 1 (1994), 45–66.
- Adam Metallo, Vincent Rossi, Jonathan Blundell, Günter Waibel, Paul Graham, Graham Fyffe, Xueming Yu, and Paul Debevec. 2015. Scanning and printing a 3D portrait of President Barack Obama. In *SIGGRAPH 2015: Studio*. ACM, 1–1.
- 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 GENEVA Leaderboard—an Extended, Living Benchmark for Evaluating and Advancing Conversational Motion Synthesis. *arXiv preprint arXiv:2410.06327* (2024).
- Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. 2008. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics (TOG)* 27, 1 (2008), 1–24.
- Shenhan Qian, Zhi Tu, Yihao Zhi, Wen Liu, and Shenghua Gao. 2021. Speech drives templates: Co-speech gesture synthesis with learned templates. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11077–11086.
- Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. 2021. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics* 9 (2021), 53–68.
- Michael Saxon, Ayush Tripathi, Yishan Jiao, Julie M Liss, and Visar Berisha. 2020. Robust estimation of hypernasality in dysarthria with acoustic model likelihood features. *IEEE/ACM transactions on audio, speech, and language processing* 28 (2020), 2511–2522.
- Thomas A Sebeok and Jean Umiker-Sebeok. 2011. *Advances in visual semiotics: The semiotic web 1992–93*. Vol. 118. Walter de Gruyter.
- Yang Song. 2021. Score-based Generative Modeling: A Brief Introduction. <https://yang-song.net/blog/2021/score/> Accessed: 2024-11-18.
- Sebastian Starke, Ian Mason, and Taku Komura. 2022. Deepphase: Periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–13.
- 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).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- Rebecca A Webb. 1997. *Linguistic features of metaphoric gestures*. University of Rochester.
- Bowen Wu, Chaoran Liu, Carlos T Ishii, and Hiroshi Ishiguro. 2021. Probabilistic human-like gesture synthesis from speech using GRU-based WGAN. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*. 194–201.
- Jing Xu, Wei Zhang, Yalong Bai, Qibin Sun, and Tao Mei. 2022. Freeform body motion generation from speech. *arXiv preprint arXiv:2203.02291* (2022).
- Sicheng Yang, Zhiyong Wu, Minglei Li, ZhenSong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. 2023a. Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. *arXiv preprint arXiv:2305.04919* (2023).
- Sicheng Yang, Zhiyong Wu, Minglei Li, Mengchen Zhao, Jiuxin Lin, Liyang Chen, and Weihong Bao. 2022. The ReprGesture entry to the GENEVA Challenge 2022. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 758–763.

- Sicheng Yang, Zunnan Xu, Haiwei Xue, Yongkang Cheng, Shaoli Huang, Mingming Gong, and Zhiyong Wu. 2024. Freetalker: Controllable speech and text-driven gesture generation based on diffusion models for enhanced speaker naturalness. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7945–7949.
- Sicheng Yang, Haiwei Xue, Zhensong Zhang, Minglei Li, Zhiyong Wu, Xiaofei Wu, Songcen Xu, and Zonghong Dai. 2023b. The DiffuseStyleGesture+ entry to the GENEVA Challenge 2023. In *Proceedings of the 2023 International Conference on Multimodal Interaction*.
- Yanzhe Yang, Jimei Yang, and Jessica Hodgins. 2020. Statistics-based motion synthesis for social conversations. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 201–212.
- Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geohyuk 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.
- Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2022. The GENEVA Challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 736–747.
- Mingyan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2022. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001* (2022).
- Chi Zhou, Tengyu Bian, and Kang Chen. 2022. Gesturermaster: Graph-based speech-driven gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 764–770.

A BVH Data Processing Pipeline

A.1 Skeleton Structure of a Character

Some skeleton joint names from the 75 motion skeletons include:

Hips, Spine, Neck, Head, RightShoulder, RightArm, RightForeArm, RightHand, LeftShoulder, LeftArm, LeftForeArm, LeftHand, RightUpLeg, RightLeg, RightFoot, RightToeBase, LeftUpLeg, LeftLeg, LeftFoot, LeftToeBase, ...

A.2 Structure of a BVH File

A BVH (Biovision Hierarchy) file is a data format that contains information about the skeleton structure and motion data of bones in a skeletal system. A BVH file consists of two main parts: the skeleton hierarchy declaration and the bone motion data.

• HIERARCHY:

- Defines the components and names of the skeleton joints, as well as the initial positions of the joints in the T-pose (motion capture actors extend their arms horizontally to form a "T").
- Defines the parent-child relationships from the root node to the leaf nodes of the skeleton, typically with the root node being the spine (Spine).
- Specifies the data to be recorded such as position or rotation angles along X , Y , Z axes of each joint over time.
- **MOTION:** A sequence of movements frame by frame, where each frame contains bone movement data as defined in the HIERARCHY section (e.g., rotation angles or positions).

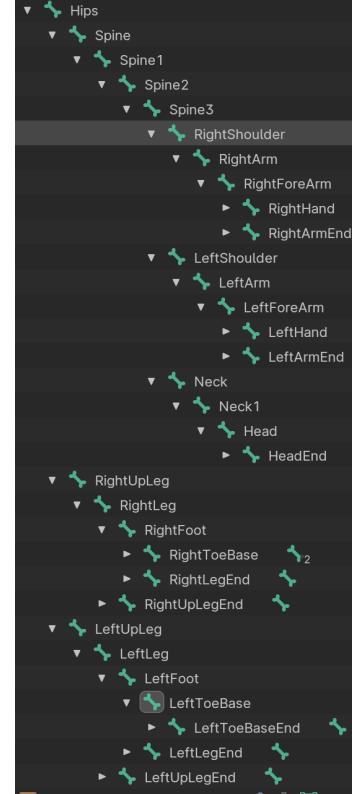
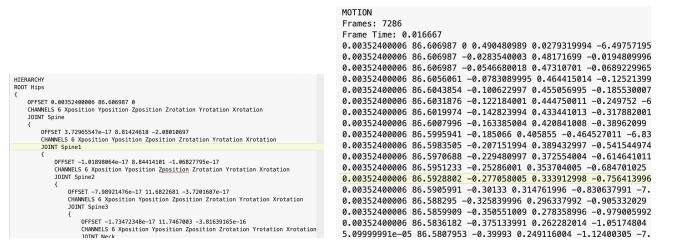


Figure 12: Character skeleton



(a) HIERARCHY in BVH file

(b) MOTION in BVH file

$\text{rotation}_i^{\text{local}} = \{\alpha, \beta, \gamma\}$ represents the rotation angles around the Z , Y , and X axes, respectively. The combined rotation in Euler space is:

$$R = R_Z(\alpha)R_Y(\beta)R_X(\gamma) \quad (11)$$

Where:

1. Rotation matrix around axis Z :

$$R_Z(\alpha) = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) & 0 \\ \sin(\alpha) & \cos(\alpha) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

2. Rotation matrix around axis Y :

$$R_Y(\beta) = \begin{bmatrix} \cos(\beta) & 0 & \sin(\beta) \\ 0 & 1 & 0 \\ -\sin(\beta) & 0 & \cos(\beta) \end{bmatrix}$$

3. Rotation matrix around axis X:

$$R_X(\gamma) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\gamma) & -\sin(\gamma) \\ 0 & \sin(\gamma) & \cos(\gamma) \end{bmatrix}$$

To compute the motion coordinates of a character, the following operation is applied:

$$\text{position}_{\text{global}} = R \cdot \text{position}_{\text{local}} + \mathbf{t} \quad (12)$$

A.3 Conversion from Euler Angles to Quaternions

To avoid Gimbal lock, Euler angle data must be converted into quaternion representation. Each bone's rotation from Euler angles in the ZYX order is represented as a quaternion $q = (q_w, q_x, q_y, q_z)$, with components calculated as follows:

First, compute the cos and sin values of half the rotation angles for each axis:

- $c_\alpha = \cos\left(\frac{\alpha}{2}\right), \quad s_\alpha = \sin\left(\frac{\alpha}{2}\right)$
- $c_\beta = \cos\left(\frac{\beta}{2}\right), \quad s_\beta = \sin\left(\frac{\beta}{2}\right)$
- $c_\gamma = \cos\left(\frac{\gamma}{2}\right), \quad s_\gamma = \sin\left(\frac{\gamma}{2}\right)$

Based on the values above, the quaternion components are computed as:

- $q_w = c_\alpha c_\beta c_\gamma + s_\alpha s_\beta s_\gamma$
- $q_x = c_\alpha c_\beta s_\gamma - s_\alpha s_\beta c_\gamma$
- $q_y = c_\alpha s_\beta c_\gamma + s_\alpha c_\beta s_\gamma$
- $q_z = s_\alpha c_\beta c_\gamma - c_\alpha s_\beta s_\gamma$

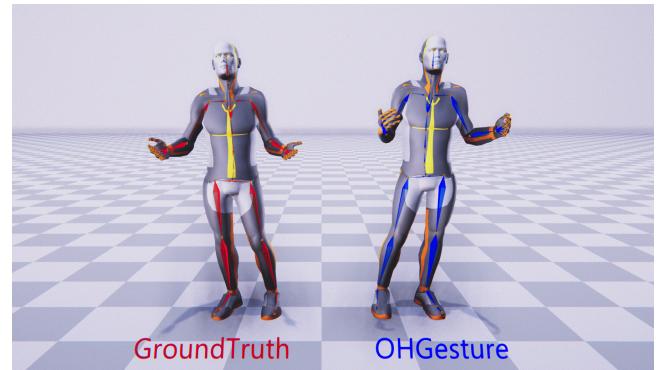
With the computed quaternion q , the global position of the bone $\mathbf{p}_{\text{global}}$ is determined by rotating the local position $\mathbf{p}_{\text{local}}$ using the formula:

$$\mathbf{p}_{\text{global}} = q \cdot \mathbf{p}_{\text{local}} \cdot q^{-1} + \mathbf{t} \quad (13)$$

where \mathbf{t} is the origin position of the bone in global space.

B Illustration of Gesture Inference Results

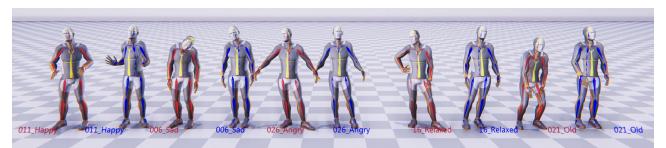
B.1 Comparison between Ground Truth Gestures and Predicted Gestures



Click the image to watch the video

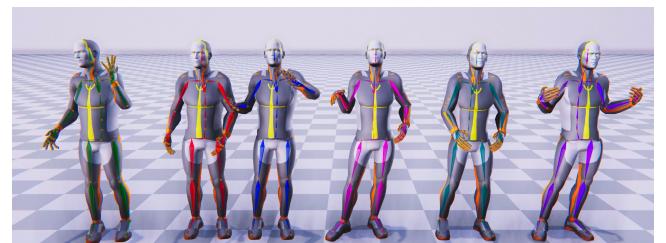
The results show both the ground truth gestures and the model's predicted gestures at frame 3821, inferred from the speech sample 003_Neutral_2_x_1_0.

B.2 Illustration of Different Emotions in Gestures



Click the image to watch the video

Generated or inferred results from the OHGesture model with different emotions. Red indicates the ground truth from the ZeroEGGS dataset, while blue shows the output generated by the OHGesture model.



Click the image to watch the video

B.3 Illustration of Gesture Generation with Out-of-Training Speech



[Click the image to watch the video](#)

Illustration of gesture generation corresponding to Steve Jobs' speech.



[Click the image to watch the video](#)

Illustration of gesture generation with synthesized speech from Microsoft Azure introducing the topic.

B.4 Illustration of Character Motion



[Click the image to watch the video](#)

Illustration of gestures extracted from 6 frames before and 6 frames after the target gesture.