Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)
Special Track on AI, the Arts and Creativity

DiffuseStyleGesture: Stylized Audio-Driven Co-Speech Gesture Generation with Diffusion Models

Sicheng Yang[1], **Zhiyong Wu**[1,4*], **Minglei Li**[2], **Zhensong Zhang**[3], **Lei Hao**[3], **Weihong Bao**[1], **Ming Cheng**[1], **Long Xiao**[1]
[1]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2]Huawei Cloud Computing Technologies Co., Ltd, Shenzhen, China
[3]Huawei Noah's Ark Lab, Shenzhen, China
[4]The Chinese University of Hong Kong, Hong Kong SAR, China
yangsc21@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn,
{liminglei29, zhangzhensong}@huawei.com

Abstract

The art of communication beyond speech there are gestures. The automatic co-speech gesture generation draws much attention in computer animation. It is a challenging task due to the diversity of gestures and the difficulty of matching the rhythm and semantics of the gesture to the corresponding speech. To address these problems, we present **DiffuseStyleGesture**, a diffusion model-based speech-driven gesture generation approach. It generates high-quality, speech-matched, stylized, and diverse co-speech gestures based on given speeches of arbitrary length. Specifically, we introduce cross-local attention and self-attention to the gesture diffusion pipeline to generate better speech-matched and realistic gestures. We then train our model with classifier-free guidance to control the gesture style by interpolation or extrapolation. Additionally, we improve the diversity of generated gestures with different initial gestures and noise. Extensive experiments show that our method outperforms recent approaches on speech-driven gesture generation. Our code, pre-trained models, and demos are available at https://github.com/YoungSeng/DiffuseStyleGesture.

1 Introduction

Body gestures and facial expressions are important tools for conveying information in human communication [Kucherenko *et al.*, 2021]. Automated generation of co-speech gestures is a crucial technology for developing lifelike avatars in movies, gaming, virtual social environments, and interactions with social robots [Nyatsanga *et al.*, 2023]. The most important issues of co-speech gesture generation are 1) how to generate gestures matching the rhythm of audio and semantics of text; and 2) how to generate diverse and stylized gestures. Recent gesture generation methods can directly generate human gestures conditioned on neutral speech [Nyatsanga *et al.*, 2023; Yoon *et al.*, 2022; Kucherenko *et al.*, 2021]. However, all these approaches still limit the learned distribution since they mainly employ GAN-based [Yoon *et al.*, 2020], VAEs [Li *et al.*, 2021a] or Flows [Alexanderson *et al.*, 2020a]. GAN-based synthesis methods suffer from mode collapse, which leads to low-quality synthesis, especially with data unseen in the training data. Methods using VAEs and Flows require a trade-off between generation quality and diversity [Tevet *et al.*, 2022; Dabral *et al.*, 2022].

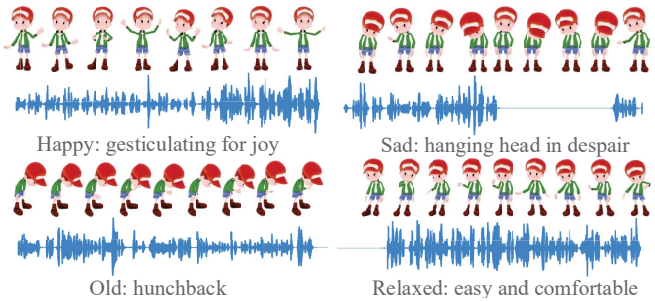


Figure 1: **Gesture examples generated by our proposed method** on various types of speech and styles. All characters used in the paper are publicly available.

Recently, diffusion models [Ho *et al.*, 2020] which are generative approaches have achieved impressive results in other domains due to their high quality and diversity of generation, such as image generation [Ramesh *et al.*, 2022], video generation [Mei and Patel, 2022], and text generation [Lin *et al.*, 2022]. These works demonstrate the ability of denoising-diffusion-based models to learn real data distributions while also providing diverse sampling and manipulation, such as editing and interpolation. However, these works do not model timing-dependent sequences to solve temporal aligned problems like speech-driven gestures, and they are computationally resource-intensive.

To generate high-quality, speech-matched, stylized, and diverse co-speech gestures, inspired by the recent progress of the denoising-diffusion-based generation, we introduce **DiffuseStyleGesture**, a versatile, controllable, and time-aware denoising-diffusion-based model for audio-driven co-speech gesture generation. Examples of the generated gesture are

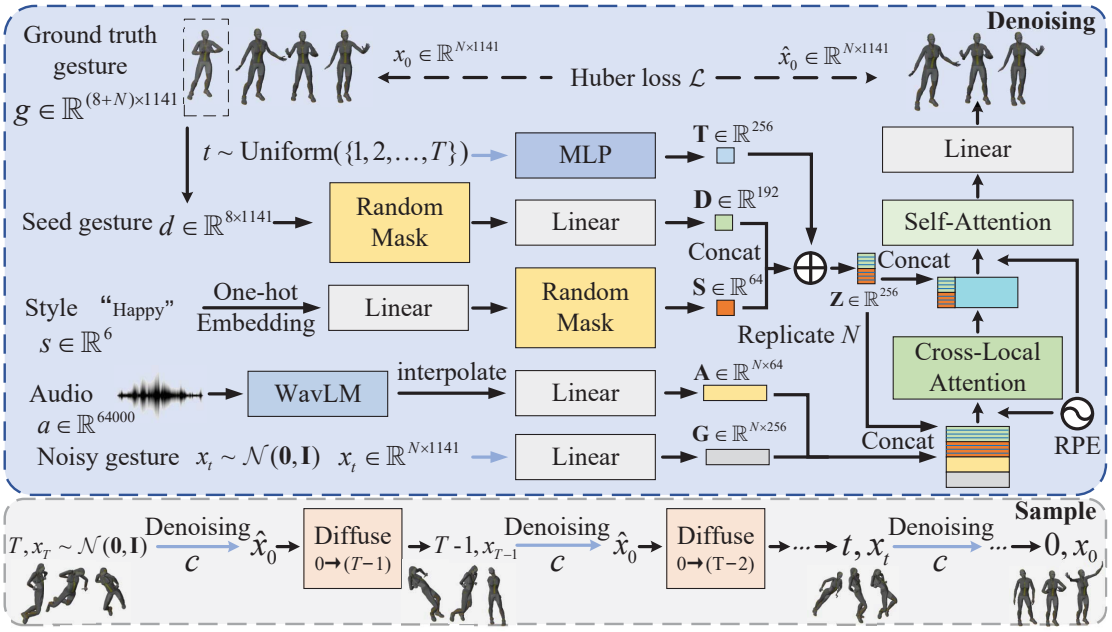


Figure 2: **(Top) Denoising module of DiffuseStyleGesture.** A noising step t and a noisy gesture sequence x_t at this noising step conditioning on c (including seed gesture d, style s, and audio a) are fed into the model. Cross-local attention and self-attention can better capture the correlations between speech and gesture based on WavLM features. Random masks in the seed gesture and style feature processing pipeline help classifier-free guidance training of the model and perform interpolation or extrapolation to achieve a high degree of control over the generated gestures. **(Bottom) Sample module of DiffuseStyleGesture.** At each step t, we predict the $\hat{x}_0$ with the denoising process based on the corresponding conditions, then add the noise to the noising step x_{t-1} with the diffuse process. This process is repeated from $t = T$ until $t = 0$.

shown in Figure 1. And the overview of our method is shown in Figure 2. We use an attention-based architecture to capture the temporal information between speech and gestures. And we find that it is better to train the model to predict the signal itself [Ramesh *et al.*, 2022; Tevet *et al.*, 2022] than to predict the noise. To align the generated gestures better with the speech, we also propose an approach that uses cross-local attention to capture local information of gestures and speech, and then uses self-attention to capture global information for better co-speech gesture generation of arbitrary length depending on the speech duration. Furthermore, we exploit WavLM features [Chen *et al.*, 2022] to consider semantic, emotional, and other information in audio to improve the generalization and robustness of our model. Finally, we use random masks to perform classifier-free guidance [Ho and Salimans, 2022] at training time and thus achieve the interpolation and editing of the control conditions. The main contributions of our work are:

- We extend the diffusion model with temporal information for audio-driven co-speech gesture generation. By virtue of the diffusion model, we can have a high degree of control over the generated gestures, e.g., editing the style of the gestures, setting the initial gestures, and generating diverse gestures.
- We use cross-local attention and global self-attention to capture feature information and make generated gestures that are more appropriate for speech.
- Extensive experiments show that our model can generate human-like, speech-matched, style-matched gestures that significantly outperform existing gesture generation methods.

2 Related Work

2.1 Co-speech Gesture Generation

Gesture generation is a complex task that requires understanding speech, gestures, and their relationships. Data-driven approaches attempt to learn gesticulation skills from human demonstrations. Present studies mainly consider four modalities: text [Alexanderson *et al.*, 2021; Yoon *et al.*, 2019], audio [Habibie *et al.*, 2021; Li *et al.*, 2021b; Ginosar *et al.*, 2019], gesture motion, and speaker identity [Yoon *et al.*, 2020; Liu *et al.*, 2022; Alexanderson *et al.*, 2020a]. [Habibie *et al.*, 2021] propose the first approach to jointly synthesize both the synchronous 3D conversational body and hand gestures, as well as 3D face and head animations. [Yi *et al.*, 2022] employ an autoencoder for face motions, and a compositional vector-quantized variational autoencoder (VQ-VAE) to generate more diverse gestures. [Xie *et al.*, 2022] introduce a VQ-VAE model to represent a pose sequence as a sequence of latent codes and develop a diffusion architecture for Text-to-Sign pose sequences generation. As for learning individual styles [Li *et al.*, 2021a; Liang *et al.*, 2022], [Yoon *et al.*, 2020] propose the first end-to-end method for generating co-speech gestures using the

tri-modality of text, audio and speaker identity. [Ahuja *et al.*, 2020] train a single model for multiple speakers while learning the style embeddings for the gestures of each speaker. [Liang *et al.*, 2022] propose a semantic energized generation method for semantic-aware gesture generation. [Ao *et al.*, 2022] disentangle both low-level and high-level embeddings of speech and motion based on linguistic theory.

Some works use motion matching methods to generate co-speech gestures [Habibie *et al.*, 2022; Zhou *et al.*, 2022]. The approach requires careful design of the database, which is directly related to the performance of the generated gestures. The length of matching needs to be balanced between quality and diversity. Furthermore, the approach also requires complex and time-consuming manual design of the matching rules.

Recently, a high-quality 3D gestures dataset ZeroEGGS [Ghorbani *et al.*, 2022] is built from multi-camera videos with style labels. This dataset is used in our work.

2.2 Diffusion Models for Motion Generation

Diffusion models excel at modeling complicated data distribution and generating vivid motion sequences. Many works integrate diffusion-based generative models into the motion domain and carefully adapt the network structure of classifier-free diffusion generative models for the human motion domain, such as based on Transformer [Kim *et al.*, 2022; Tevet *et al.*, 2022; Ren *et al.*, 2022; Zhou and Wang, 2022]. [Chang *et al.*, 2022b] design a multi-task architecture of diffusion model and use adversarial and physical regulations for human motion synthesis. [Dabral *et al.*, 2022] introduce MoFusion to generate long, temporally plausible, and semantically accurate motions. A physics-guided motion diffusion model [Yuan *et al.*, 2022] incorporates physical constraints into the diffusion process.

[Ginosar *et al.*, 2019] propose a cross-modal translation method based on the speech-driven gestures of a single speaker. [Li *et al.*, 2022] propose a cross-conditional causal attention layer to keep the coherence of the generated body. [Chang *et al.*, 2022a] use locality-constraint attention mechanism and achieve the best gesture-speech appropriateness in the full-body level of the GENEA 2022 gesture generation challenge [Yoon *et al.*, 2022]. MotionDiffuse [Zhang *et al.*, 2022] using cross attention enables probabilistic mapping, realistic synthesis, and multi-level manipulation. In our work, we use cross-local attention to capture local information of gestures and speech; and then use self-attention to capture global information to make the generated gestures match better with the speech.

3 Our Approach

3.1 Diffusion Model for Gesture Generation

Our idea is to generate gestures with a diffusion model [Ho *et al.*, 2020] by learning to gradually denoising starting from pure noise. As shown in Figure 2, the diffusion model consists of two parts: the forward process (diffusion process) q and the reverse process (denoising process) p_θ.

Diffusion Process. The diffusion process q is modeled as a Markov noising process. We denote the generated gesture as x, which has the same dimension as an observation data $x_0 \sim q(x_0)$, $q(x_0)$ denotes the distribution of the real data. According to a variance schedule $\beta_1, \beta_2, \ldots, \beta_T$ ($0 < \beta_1 < \beta_2 < \cdots < \beta_T < 1$, T is the total time step), we add Gaussian noise

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}\right) \tag{1}$$

to the gesture at each time t gradually, and if the schedule is properly designed and T is large, pure noise

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}) \tag{2}$$

will be obtained at the end.

Denoising Process. The denoising process p_θ is a process of learning parameter θ via a neural network. Assuming that the denoising process also conforms to a Gaussian distribution, i.e., the noise x_t at time t is used to learn μ_θ, Σ_θ, then

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{3}$$

For calculation convenience, we assume that $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i$. Then the noisy gesture x_t at time t can be written as

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}\right) \tag{4}$$

The network is optimized by minimizing the difference between the real noise ϵ and the predicted noise $\epsilon_\theta(x_t, t)$ [Ho *et al.*, 2020]. When sampling, we can learn the mean $\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right)$ by fix the variance.

Framework. Our goal is to synthesize a human gesture $x^{1:N}$ of length N given conditions c. In our work, we follow [Ramesh *et al.*, 2022; Tevet *et al.*, 2022] to predict the signal itself instead of predicting $\epsilon_\theta(x_t, t)$ [Ho *et al.*, 2020]. The Denoising module reconstructs the original signal x_0 based on the input noise x_t, noising step t and conditions c

$$\hat{x}_0 = \text{Denoise}(x_t, t, c) \tag{5}$$

Then the Denoising module can be trained by optimizing the Huber loss [Huber, 1992] between the generated poses $\hat{x}_0$ and the ground truth human gestures x_0 on the training examples:

$$\mathcal{L} = E_{x_0 \sim q(x_0|c), t\sim[1,T]}\left[\text{HuberLoss}(x_0 - \hat{x}_0)\right] \tag{6}$$

3.2 Attention-based Speech-driven Gesture Generation Model

Feature Processing in Denoising module. As shown in Figure 2, gestures are generated based on noising step t, noisy gesture x_t and conditions c (including audio a, style s, and seed gesture d). For each feature, the processing pipeline is as follows:

- Noising step: During training, noising step t is sampled from a uniform distribution of $\{1, 2, \ldots, T\}$, with the same position encoding as [Vaswani *et al.*, 2017], and then mapped to a space $\mathbf{T}$ of dimension 256 by a multilayer perceptron (MLP).

- Noisy gesture: During training, x_t is the noisy gesture with the same dimension as the real gesture x_0 obtained by sampling from the standard normal distribution $\mathcal{N}(0, \mathbf{I})$. When sampling, the initial noisy gesture x_T is sampled from the standard normal distribution and the other $x_t, t < T$ is the result of the previous noising step. Then the dimension is adjusted to 256 as $\mathbf{G}$ by a linear layer.
- Audio: All audio recordings are downsampled to 16kHz, and features are generated from the pre-trained models of WavLM Large [Chen *et al.*, 2022]. We use linear interpolation to align WavLM features and gesture x_0 in the time dimension to 20fps, and then use a linear layer to reduce the dimension of features to 64 forming the final audio feature sequence $\mathbf{A}$.
- Style: The styles of gestures are represented as one-hot vectors where only one element of a selected style is nonzero, mapping to the 64-dimensional space $\mathbf{S}$ via a linear layer.
- Seed gesture: Seed gesture helps to make smooth transitions between consecutive syntheses [Yoon *et al.*, 2020]. Please see our supplementary material for more detail regarding the ground truth gestures clip $g \in \mathbb{R}^{(8+N)\times 1141}$. The first 8 frames of the gestures clip g are used as the seed gesture d and the remaining N frames are used as the real gesture x_0 to calculate loss $\mathcal{L}$. Then we map the feature dimensions of seed gesture d to the space $\mathbf{D}$ of 192 dimensions using a linear layer. The length of our generated gesture is 4 seconds and we resample the gesture animation to 20 fps, then N = 80.

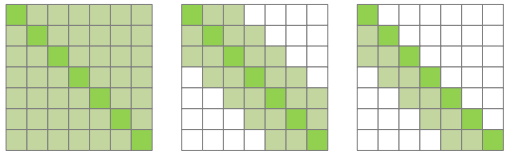
Model in Denoising module. We implement denoising with the attention-base architecture. Before utilizing long-range correlations, it is advisable to build up representations with local context [Rae and Razavi, 2020].

We concatenate the seed gesture $\mathbf{D}$ and style $\mathbf{S}$ together to form a 256-dimensional vector, and then add the information of the noising step $\mathbf{T}$ to form $\mathbf{Z}$. Our network takes the vector $\mathbf{Z}$ and stacks its replicates into a sequence feature to align with the timeline of audio and gesture features, which are then concatenated with the audio $\mathbf{A}$ and gesture $\mathbf{G}$ as the input to the cross-local attention network. Our proposed cross-local attention for co-speech gesture generation is based on Routing Transformer [Roy *et al.*, 2021], which shows that local attention is important in building intermediate representations, as shown in Figure 3(c).

After that, we concatenate the output of cross-local attention with $\mathbf{Z}$ and feed it into the self-attention network, as shown in Figure 3(a). The self-attention mechanism is similar to Transformer [Vaswani *et al.*, 2017] encoder, which determines the computational dependencies between the sequential elements of the data, and is implemented as

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{M}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T + \mathbf{M}}{\sqrt{C}}\right)\mathbf{V} \tag{7}$$

where $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ denote the query, key and value from input, and $\mathbf{M}$ is the mask, which determines the type of attention patterns. We use the same relative position encoding (RPE) mechanism as [Kitaev *et al.*, 2020] so that the temporal effect on gesture translation is invariant. Finally, the output of self-attention is mapped back to the same dimension as x_0 after a linear layer.



(a) full self-attention (b) sliding window attention (c) cross-local attention

Figure 3: Different patterns of attention used in our experiments, where (a) and (c) are attention mechanisms used in our model and (b) is a pattern compared in Section 4.3. The rows represent the outputs and the columns represent the inputs. The colored squares highlight the relevant elements for each row of output.

From Figure 3, we can find these different attention mechanisms can be achieved by simply adjusting the corresponding mask $\mathbf{M}$, we also experimented with sliding window attention in Longformer [Beltagy *et al.*, 2020], the results are analyzed in Section 4.3.

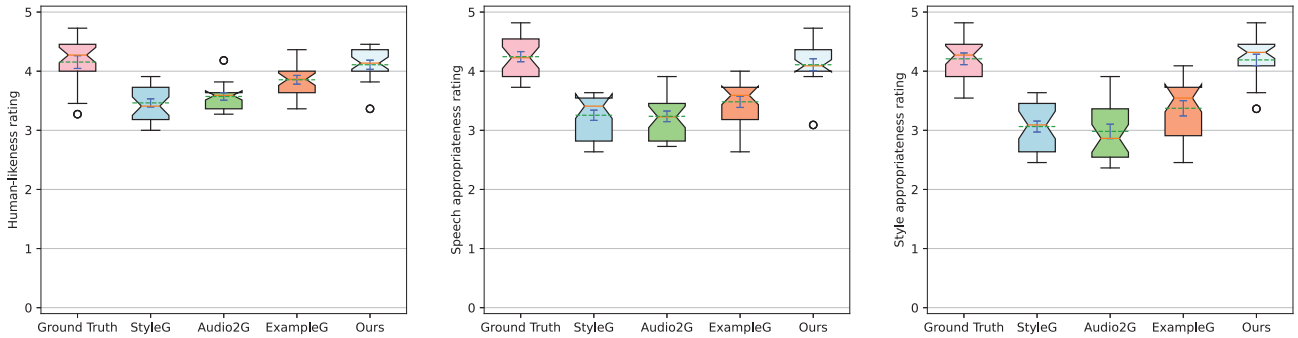
Sample Module. The final co-speech gesture is given by splicing a number of clips of length N. The seed gesture for the first clip can be generated by randomly sampling a gesture from the dataset or by setting it to the average gesture. Then the seed gesture for other clips is the last 8 frames of the gesture generated in the previous clip. For every clip, in every noising step t, we predict the clean gesture $\hat{x}_0$ =Denoise(x_t, t, c), and add the noise to the noising step x_{t-1} using Equation (1) with the diffuse process. This process is repeated from $t = T$ until x_0 is reached (Figure 2 bottom).

3.3 Style-controllable Gesture Generation

Since the algorithm generates gestures based on control conditions, the control conditions can be not only audio a but also style s, or seed gesture d, etc. As shown in Figure 2, we refer to [Ho and Salimans, 2022; Tevet *et al.*, 2022] and add random masks to the pipeline of seed gesture d and style s feature processing for classifier-free learning, which enables accurate control of different conditions. Then, the classifier-free guidance of gestures generation can be achieved by combining the predictions of the conditional model $\text{Denoise}\left(x_t, t, c_1\right), c_1 = [d, s, a]$ and the unconditional model $\text{Denoise}\left(x_t, t, c_2\right), c_2 = [\varnothing, \varnothing, a]$ during the training process, as Equation (8).

$$\hat{x}_{0_{\gamma,c_1,c_2}} = \gamma\,\text{Denoise}\left(x_t, t, c_1\right) + (1-\gamma)\,\text{Denoise}\left(x_t, t, c_2\right) \tag{8}$$

In practice, the Denoising module learns both the conditioned and the unconditioned distributions by randomly masking 10% of the samples using Bernoulli masks. Then, as for style s in condition, we can generate style-controlled gestures when sampling by interpolating or even extrapolating the two variants using γ, as $c_1 = [d, s_1, a], c_2 = [d, s_2, a]$ in Equation (8). Please refer to our supplementary material for training details such as dataset and implementation details.



(a) Box plot of ratings in human-likeness.

(b) Box plot of ratings in gesture-speech appropriateness.

(c) Box plot of ratings in gesture-style appropriateness.

Figure 4: Box plot visualizing comparison results of MOS for different models in different dimensions. The box extends from the first lower quartile (Q1) to the third greater quartile (Q3) of the data. The red line denotes the median. The notches represent the 95% confidence interval (CI) around the median. When the CI is less than Q1 or greater than Q3, the notch extends beyond the box, giving it a unique "flipped" appearance. We have also marked the mean and its 95% CI in the figure with a green dashed line and a blue vertical line, respectively.

4 Experiments

4.1 Comparison to Existing Methods

We compare our proposed model with StyleGestures [Alexanderson *et al.*, 2020b], Audio2Gestures [Li *et al.*, 2021b], ExampleGestures [Ghorbani *et al.*, 2022]. Currently, speech-driven gestures lack objective metrics that are consistent with human subjective perception [Yoon *et al.*, 2022; Kucherenko *et al.*, 2021; Alexanderson *et al.*, 2022], even for Fréchet gesture distance (FGD) [Yoon *et al.*, 2020; Dabral *et al.*, 2022], so all our experimental scoring are done by human subjective evaluation. We conduct the evaluation on three dimensions. The first two follow the evaluation in GENEA [Yoon *et al.*, 2022], which evaluates human-likeness and gesture-speech appropriateness. The third dimension is gesture-style appropriateness.

User Study. To understand the real visual performance of our method, we conduct a user study among the gesture sequences generated by each compared method and the ground truth motion capture data. The length of the evaluated clips ranged from 11 to 51 seconds, with an average length of 31.6 seconds. Note that the clip gestures used for the subjective evaluation here are longer compared to the GENEA [Yoon *et al.*, 2022] evaluation (8-10 seconds), as a longer period time could produce more pronounced and convincing appropriateness results [Yang *et al.*, 2022]. Participants rated at a 1-point interval from 5 to 1, with labels (from best to worst) of "excellent", "good", "fair", "poor", and "bad". More details about the user study are shown in the supplementary material.

The mean opinion scores (MOS) on human-likeness, speech appropriateness, and style appropriateness are reported in Figure 4. If the notches of the two boxes do not overlap, we can consider this as strong evidence that the distributions are significantly different [McGill *et al.*, 1978]. Our method significantly surpasses the compared state-of-the-art methods with human-likeness, gesture-speech, and gesture-style appropriateness, and even produces competitive results with ground truth in all three dimensions. According to the feedback from participants, our generated gestures are "more semantically relevant", "more natural", and "match the style", while our approach has "foot-sliding" compared to Ground Truth. However, this is a common problem for non-physical-based motion generation systems and could be solved by post-processing [Ghorbani *et al.*, 2022; Luvizon *et al.*, 2023].

4.2 Gesture Controllability

Style Control. Assuming that the neutral audio does not affect the gesture style, then we can generate stylized gestures with a neutral speech by setting $\gamma = 1$ and s in equation (8). We choose two speech segments in the test set with neutral audio to generate six stylized gestures respectively. Figure 5 illustrates generated gesture $\hat{x}_t$ of different input style s with the same neutral audio visualized by the tSNE method.

We also plotted the body skeleton generated by the corresponding style in the figure, and it can be noticed that for the 'old' style, its waist and knees are more bent, and its hands are basically on the knees or waist; for the 'sad' style, its head is hanging and its hands are in a lower position; for the 'relaxed' style, its hips are forward and its standing posture is relaxed; for the 'angry' style, its hands move up and down quickly. Note that although the differences between the 'neutral' and 'happy' style gestures are still relatively obvious in the stylistic visualization of the skeleton, i.e., for the 'happy' style, its hand position is higher and its amplitude is larger, their tSNE is almost coupled together. In our analysis, this happens because the so-called neutral speech still contains information such as emotions and semantics that is contained in the WavLM features. The balance of the style from audio a and from style s can be further controlled by the editing the style intensity γ.

Style Edit. To further analyze the relationship between style intensity and the style implied by the speech, we choose the 'happy' style and the 'old' style and set γ = 1 and 3 in equation 8. Also, to compare the results, we set γ = 0.5 in equation 8, in order to interpolate the different styles. The

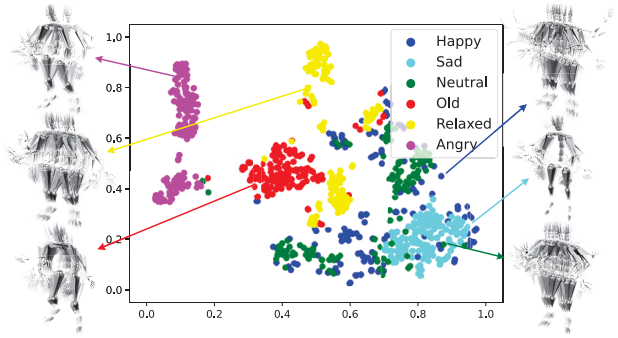


Figure 5: The tSNE visualization of gestures with different styles and the shadow maps of the skeletal gesture with the corresponding style. For example, for the 'old' gesture, its waist and knees are more bent, and its hands are basically on the knees or waist.

other parameters remain unchanged, and we plot a 12-second, FPS = 1 gesture generation result as shown in Figure 6.

As shown in Figure 6, we can see that when we use the 'happy' style with $\gamma = 3$, both its body rotation and hand movements are the largest, and its hand positions are the highest; in contrast, when we use the 'old' style with $\gamma = 3$, its waist is the most bent, the hands are barely lifted, and there is no much change in the whole movement sequence; As for the other three results, the intensities of their styles are in the middle of the above two, and the style gradually changes from happy to old from top to bottom. Due to our model architecture, the generated gestures and speech are more appropriate, even though these styles are not the same. Notice that when we use 'happy' and 'old' styles with $\gamma = 0.5$, the result is closer to using the 'happy' style with $\gamma = 1$, while the 'old' style is almost imperceptible. This observation further validates the previous finding that the 'happy' style is embedded in the 'neutral' speech used for testing. This finding is useful, for example, we found in our experiments that if we want to control the 'happy' speech to generate the 'sad' gesture, $\gamma = 1$ is basically ineffective because the model can learn the happy style from the speech. Since there is such a coupling between the style of speech and the style of gestures, then setting a larger γ can edit the style better. Thus we are able to generate gestures that do not exist in the original dataset (e.g., gestures for 'happy' speech in the 'sad' style) by style intensity.

User Study. Further, we would like to explore the relationship between style intensity and human-likeness and speech appropriateness, so we conducted a user study. To avoid styles in speech from influencing participants' scoring, as before, we only control the intensity of the styles for a neutral speech and then asked participants to score the three previous dimensions. ExampleGesture [Ghorbani *et al.*, 2022] can also control the generation of different styles of gestures from the same speech. Hence we choose it as the baseline model. Since the gestures generated here do not exist in the dataset, the source neutral speech with neutral style is used as a reference. The results are shown in Figure 7.

The results show that our model is similar to ExampleGesture in terms of gesture-style appropriateness of the results at $\gamma = 1$, and our human-likeness and speech appropriateness

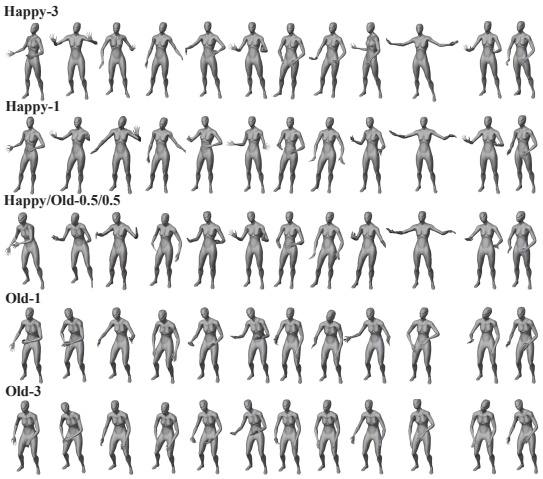


Figure 6: Style editing and interpolation results. From top to bottom, the body twists and hand movements gradually decrease and the hand position becomes lower. Despite the change in style, the generated gestures still match well with speech in different styles.

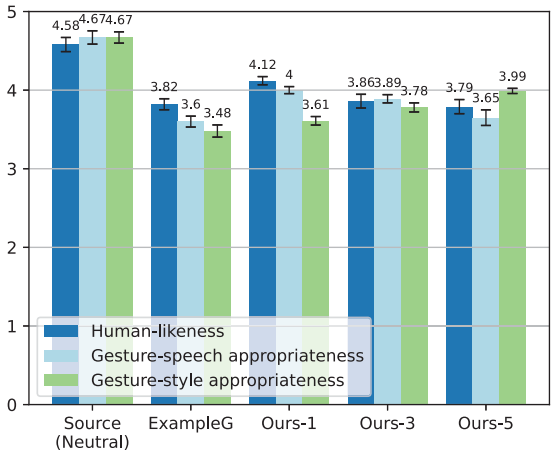


Figure 7: Average results of MOS with 95% confidence intervals for three dimensions. 'Ours-γ' denotes the style control intensity γ of our model. Our model significantly outperform ExampleGesture overall and could editing the intensity of the styles. The parameter γ increases and the other two scores will reasonably decrease.

significantly exceed ExampleGesture. Meanwhile, the style is significantly more appropriate when γ increases, but the scores of the other two dimensions decrease. This is also intuitive, i.e., if the intensity of the 'old' style is too high, the hands are barely lifted and the entire motion sequence are small in amplitude, so it looks less human-like and less appropriate to the speech. We also find that the results of generating style control (Figure 7) were degraded compared to the results of directly generating the style corresponding to the speech (Figure 4). We believe that controlling one style of speech to generate another style of gesture is in itself a "difficult and conflicting" task because the style of speech and the style of gesture are still related and coupled together.

Generate diverse gestures. Due to our model architecture,

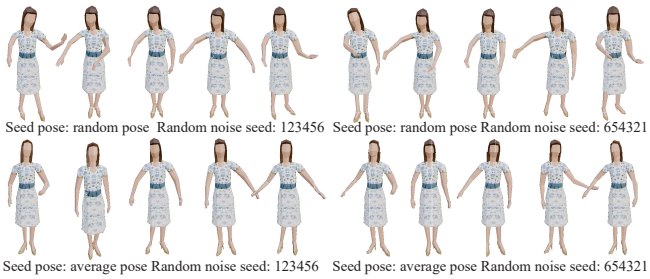


Figure 8: Visualization of the diversity of generated gestures. People make different co-speech gestures at different moments in different states. Just like real people, for the same speech, our method is able to generate different gestures with different seed gesture or with different noisy gesture.

Name	Human likeness ↑	Gesture-speech appropriateness ↑
Ground Truth	4.15 ± 0.11	4.25 ± 0.09
Ours	$\mathbf{4.11 \pm 0.08}$	$\mathbf{4.11 \pm 0.10}$
− WavLM	4.05 ± 0.10	3.91 ± 0.11
− cross-local attention	3.76 ± 0.09	3.51 ± 0.15
− self-attention	3.55 ± 0.13	3.08 ± 0.10
− attention + GRU	3.10 ± 0.11	2.98 ± 0.14
+ forward attention	3.75 ± 0.15	3.23 ± 0.24

Table 1: Ablation studies results. '−' indicates modules that are not used and '+' indicates additional modules. Bold indicates the best metric.

even for the same speech and style, different noisy gesture and different seed gesture could generate different results, as shown in Figure 8. This is the same as real human speech, which creates diverse co-speech gestures related to the initial position. Our analysis before was performed on the style dimension. Note that the model also adds a random mask to the processing of the seed gesture, so it can also interpolate and extrapolate different seed gestures to control the generation of different and diverse initial position gesture.

4.3 Ablation Studies

Moreover, we conduct ablation studies to address the performance effects of different components in our model. Since gesture-style appropriateness can be controlled by parameter and affect the other two dimensions, we set γ to 1 and score only human-likeness and speech appropriateness for ease of comparison. The results of our ablations studies are summarized in Table 1. The visual comparisons of this study can be also referred to the supplementary video. We explore the effectiveness of the following components: (1) WavLM features (2) local attention (3) local attention pattern (4) self-attention (5) attention. We conduct the experiments on each of the five components, respectively.

User Study. Supported by the results in Table 1, when we do not use the WavLM feature but use the first 13 coefficients of the Mel-frequency cepstral coefficients (MFCC) instead, the scores of both dimensions decreased, especially the speech appropriateness. This is because the features extracted by the pre-trained WavLM model contain more information such as semantics and emotions, which is helpful to generate the corresponding gestures. When there is no cross-local attention, the scores of both dimensions drop a lot. Because many gesture generation steps only involve short-range correlations, local attention can capture local information better, which is consistent with the observation of [Rae and Razavi, 2020]. Only self-attention relying on global information of long sequences becomes less effective. Both human-likeness and gesture-speech appropriateness drop more when self-attention is removed, suggesting that self-attention is more important than local attention because there is inherent asynchrony in speech and gesture, and it is difficult to learn enough gestural information from only a local window (nearly half a second) of speech. When attention is not used, we replace it with a GRU-based model, which has the worst results among all models, further illustrating the effectiveness of the attention mechanism. In addition, we experiment using the attention structure in Figure 3(b) and find that the effect gets worse. The only difference between adding forward attention and the cross-local attention used in our model is that the gesture is generated with an extra look at the speech information in a future window. This is an exciting finding, although there is an inherent asynchrony between speech and gesture, in some ways it could indicate that gestures are more related to a small period of time in the present and the past and not to a short period of time in the future. In other words, people are more likely to say "hello" before waving than to wave before saying "hello". It is also possible that different people have different styles and this dataset has only one actor that needs to be studied further.

5 Discussion and Conclusion

In this paper, we propose DiffuseStyleGesture, a diffusion model based method for audio-driven co-gesture generation. DiffuseStyleGesture demonstrates three major strengths: 1) Based on a diffusion model, probabilistic mapping enhances diversity while enabling the generation of high-quality, human-like gestures. 2) Our model synthesis gestures match the audio rhythm and text semantics based on cross-local and self-attention mechanisms. 3) Using the classifier-free guidance training approach, we can manipulate specific conditions, i.e., style and initial gesture, and perform interpolation or extrapolation to achieve a high degree of control over the generated gestures. The subjective evaluation shows that our model outperforms existing arts on the temporal task of audio-driven co-gesture generation and demonstrates superior style manipulation. There is room for improvement in this research, for example, solving the problem of many sampling steps and long-time consumption of diffusion methods for use in real-time systems is our future research direction.

Acknowledgments

This work is supported by National Natural Science Foundation of China (62076144), Shenzhen Science and Technology Program (WDZC20220816140515001, JCYJ20220818101014030) and Shenzhen Key Laboratory of next generation interactive media innovative technology (ZDSYS20210623092001004).

References

[Ahuja *et al.*, 2020] Chaitanya Ahuja, Dong Won Lee, Yukiko I. Nakano, and Louis-Philippe Morency. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *Computer Vision - ECCV*, volume 12363 of *Lecture Notes in Computer Science*, pages 248–265, 2020.

[Alexanderson *et al.*, 2020a] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. Style-controllable speech-driven gesture synthesis using normalising flows. *Comput. Graph. Forum*, 39(2):487–496, 2020.

[Alexanderson *et al.*, 2020b] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. Style-controllable speech-driven gesture synthesis using normalising flows. *Computer Graphics Forum*, 39(2):487–496, 2020.

[Alexanderson *et al.*, 2021] Simon Alexanderson, Éva Székely, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. Generating coherent spontaneous speech and gesture from text. *CoRR*, abs/2101.05684, 2021.

[Alexanderson *et al.*, 2022] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *CoRR*, abs/2211.09707, 2022.

[Ao *et al.*, 2022] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Trans. Graph.*, 41(6):209:1–209:19, 2022.

[Beltagy *et al.*, 2020] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *CoRR*, abs/2004.05150, 2020.

[Chang *et al.*, 2022a] Che-Jui Chang, Sen Zhang, and Mubbasir Kapadia. The ivi lab entry to the genea challenge 2022 – a tacotron2 based method for co-speech gesture generation with locality-constraint attention mechanism. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, ICMI '22, page 784–789, 2022.

[Chang *et al.*, 2022b] Ziyi Chang, Edmund JC Findlay, Haozheng Zhang, and Hubert PH Shum. Unifying human motion synthesis and style transfer with denoising diffusion probabilistic models. *arXiv preprint arXiv:2212.08526*, 2022.

[Chen *et al.*, 2022] 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. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 16(6):1505–1518, 2022.

[Dabral *et al.*, 2022] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. *CoRR*, abs/2212.04495, 2022.

[Ghorbani *et al.*, 2022] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F Troje, and Marc-André Carbonneau. Zeroeggs: Zero-shot example-based gesture generation from speech. *arXiv preprint arXiv:2209.07556*, 2022.

[Ginosar *et al.*, 2019] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3497–3506, 2019.

[Habibie *et al.*, 2021] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. Learning speech-driven 3d conversational gestures from video. In *IVA '21: ACM International Conference on Intelligent Virtual Agents*, pages 101–108, 2021.

[Habibie *et al.*, 2022] Ikhsanul Habibie, Mohamed Elgharib, Kripasindhu Sarkar, Ahsan Abdullah, Simbarashe Nyatsanga, Michael Neff, and Christian Theobalt. A motion matching-based framework for controllable gesture synthesis from speech. In *SIGGRAPH '22: Special Interest Group on Computer Graphics and Interactive Techniques Conference*, pages 46:1–46:9, 2022.

[Ho and Salimans, 2022] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[Huber, 1992] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. 1992.

[Kim *et al.*, 2022] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. *arXiv preprint arXiv:2209.00349*, 2022.

[Kitaev *et al.*, 2020] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *8th International Conference on Learning Representations, ICLR April 26-30*, 2020.

[Kucherenko *et al.*, 2021] Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA challenge 2020. In *IUI '21: 26th International Conference on Intelligent User Interfaces, April 13-17*, pages 11–21, 2021.

[Li *et al.*, 2021a] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV*, pages 11273–11282, 2021.

[Li *et al.*, 2021b] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *2021*

IEEE/CVF International Conference on Computer Vision (ICCV), pages 11273–11282, 2021.

[Li *et al.*, 2022] Siyao Li, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic GPT with choreographic memory. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR June 18-24*, pages 11040–11049, 2022.

[Liang *et al.*, 2022] Yuanzhi Liang, Qianyu Feng, Linchao Zhu, Li Hu, Pan Pan, and Yi Yang. SEEG: semantic energized co-speech gesture generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 10463–10472, 2022.

[Lin *et al.*, 2022] Zhenghao Lin, Yeyun Gong, Yelong Shen, Tong Wu, Zhihao Fan, Chen Lin, Weizhu Chen, and Nan Duan. GENIE: large scale pre-training for text generation with diffusion model. *CoRR*, abs/2212.11685, 2022.

[Liu *et al.*, 2022] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 10452–10462, 2022.

[Luvizon *et al.*, 2023] Diogo Luvizon, Marc Habermann, Vladislav Golyanik, Adam Kortylewski, and Christian Theobalt. Scene-aware 3d multi-human motion capture from a single camera. *CoRR*, abs/2301.05175, 2023.

[McGill *et al.*, 1978] Robert McGill, John W Tukey, and Wayne A Larsen. Variations of box plots. *The american statistician*, 32(1):12–16, 1978.

[Mei and Patel, 2022] Kangfu Mei and Vishal M. Patel. VIDM: video implicit diffusion models. *CoRR*, abs/2212.00235, 2022.

[Nyatsanga *et al.*, 2023] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. A comprehensive review of data-driven co-speech gesture generation. *arXiv preprint arXiv:2301.05339*, 2023.

[Rae and Razavi, 2020] Jack W. Rae and Ali Razavi. Do transformers need deep long-range memory? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7524–7529, 2020.

[Ramesh *et al.*, 2022] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[Ren *et al.*, 2022] Zhiyuan Ren, Zhihong Pan, Xin Zhou, and Le Kang. Diffusion motion: Generate text-guided 3d human motion by diffusion model. *arXiv preprint arXiv:2210.12315*, 2022.

[Roy *et al.*, 2021] Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Trans. Assoc. Comput. Linguistics*, 9:53–68, 2021.

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

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Xie *et al.*, 2022] Pan Xie, Qipeng Zhang, Zexian Li, Hao Tang, Yao Du, and Xiaohui Hu. Vector quantized diffusion model with codeunet for text-to-sign pose sequences generation. *arXiv preprint arXiv:2208.09141*, 2022.

[Yang *et al.*, 2022] Sicheng Yang, Zhiyong Wu, Minglei Li, Mengchen Zhao, Jiuxin Lin, Liyang Chen, and Weihong Bao. The reprgesture entry to the GENEA challenge 2022. In *International Conference on Multimodal Interaction, ICMI, November 7-11*, pages 758–763, 2022.

[Yi *et al.*, 2022] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. *arXiv preprint arXiv:2212.04420*, 2022.

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

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

[Yoon *et al.*, 2022] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. The GENEA challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *International Conference on Multimodal Interaction, ICMI, November 7-11*, pages 736–747, 2022.

[Yuan *et al.*, 2022] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. *arXiv preprint arXiv:2212.02500*, 2022.

[Zhang *et al.*, 2022] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

[Zhou and Wang, 2022] Zixiang Zhou and Baoyuan Wang. Ude: A unified driving engine for human motion generation. *arXiv preprint arXiv:2211.16016*, 2022.

[Zhou *et al.*, 2022] Chi Zhou, Tengyue Bian, and Kang Chen. Gesturemaster: Graph-based speech-driven gesture generation. In *International Conference on Multimodal Interaction, ICMI*, pages 764–770, 2022.