Electronic Research Archive

ERA, 32(9): 5392–5408.
DOI: 10.3934/era.2024250
Received: 23 June 2024
Revised: 11 September 2024
Accepted: 14 September 2024
Published: 26 September 2024

http://www.aimspress.com/journal/ERA

Research article

Audio2DiffuGesture: Generating a diverse co-speech gesture based on a diffusion model

Hongze Yao[1], Yingting Xu[1], Weitao WU[1], Huabin He[1], Wen Ren[2,*] and Zhiming Cai[1,3,*]

[1] School of Electronics, Electrical Engineering and Physics, Fujian University of Technology, Fuzhou 350118, China
[2] School of Mechanical and Electric Engineering, Sanming University, Sanming 365004, China
[3] National Demonstration Center for Experimental Electronic Information and Electrical Technology Education, Fujian University of Technology, Fuzhou 350118, China

*** Correspondence:** Email: rw@fjsmu.edu.cn, caizm@fjut.edu.cn.

Abstract: People use a combination of language and gestures to convey intentions, making the generation of natural co-speech gestures a challenging task. In audio-driven gesture generation, relying solely on features extracted from raw audio waveforms limits the model's ability to fully learn the joint distribution between audio and gestures. To address this limitation, we integrated key features from both raw audio waveforms and Mel-spectrograms. Specifically, we employed cascaded 1D convolutions to extract features from the audio waveform and a two-stage attention mechanism to capture features from the Mel-spectrogram. The fused features were then input into a Transformer with cross-dimension attention for sequence modeling, which mitigated accumulated non-autoregressive errors and reduced redundant information. We developed a diffusion model-based Audio to Diffusion Gesture (A2DG) generation pipeline capable of producing high-quality and diverse gestures. Our method demonstrated superior performance in extensive experiments compared to established baselines. Regarding the TED Gesture and TED Expressive datasets, the Fréchet Gesture Distance (FGD) performance improved by 16.8 and 56%, respectively. Additionally, a user study validated that the co-speech gestures generated by our method are more vivid and realistic.

Keywords: co-speech gesture; cross-modal; human-computer interaction; diffusion model; attention mechanism

1. Introduction

In human-human dialogue systems, particularly in scenarios such as speeches, co-speech gestures serve as a crucial means for speakers to convey their intentions through non-verbal behavior [1–3]. Psycholinguistic studies indicate that natural body movements, such as arm waving, nodding, and shaking the head, enrich the speaker's viewpoints and foster interactive communication with the audience [1,4]. With the advancements in Artificial Intelligence Generated Content (AIGC), producing natural and diverse co-speech gestures has become one of the key challenges in current generative tasks. Particularly in virtual characters for games and films, creating expressive gestures significantly enhances the experience for players and audiences [5].

Previous research on co-speech gesture generation was based on rule-based methods [6,7]. Gesture generation systems developed by meticulously designing correspondences between speech and gesture units can produce high-quality gestures. However, these gestures lack diversity and require significant manual effort. With the progress in deep learning models, co-speech gesture generation has shifted towards data-driven approaches. Figure 1 shows a schematic diagram of the synthesized gestures. Researchers [8–11] have used adversarial training [12] to synthesize gestures, achieving impressive results. However, maintaining a balance between the generator and discriminator is difficult, often resulting in unstable training and mode collapse. Diffusion models [13] have gained widespread attention for their outstanding performance in generative tasks. In the many-to-many mapping scenario of co-speech gesture generation, the diffusion model can learn and approximate complex distributions. Therefore, we employ the latent diffusion model to reduce irregular human motion in audio-driven motion synthesis, resulting in high-quality and diverse co-speech gestures.

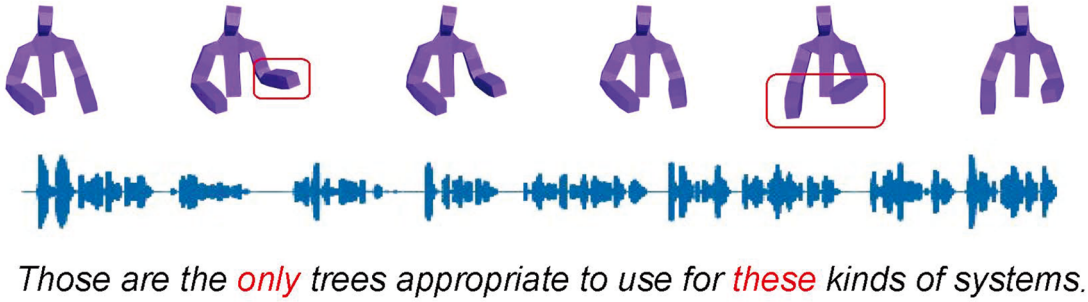


Figure 1. Visualization of synthesis co-speech gesture.

These methods [8–10] utilize multimodal inputs, such as audio, text, and speaker identity, to train generative models for synthesizing gestures. However, they have not fully explored the impact of useful audio information on gesture synthesis. Both raw audio waveforms and Mel-spectrograms contain rich audio information. Previous work [8,14–16] has extracted features from raw audio waveforms only through the decoder's final layer. In contrast, we combine the Mel-spectrogram with the raw audio waveform to achieve more detailed feature extraction across the audio space. In synthesizing long-sequence gestures, early RNN-based models [11,17,18] tend to accumulate errors over time, leading to repetitive and stagnant gestures. Transformer-based models, however, can effectively capture long-term dependencies using positional encoding to retain the sequence order of

the input data. Consequently, we utilize a Transformer model [19] to process the fused multimodal data. Additionally, we introduce a cross-dimension attention mechanism to mitigate the redundancy arising from concatenating features of the two audio modalities.

Our major contributions are as follows:

1) Using latent diffusion concepts, we establish a powerful Audio to Diffusion Gesture (A2DG) generation pipeline that synthesizes gestures with high quality and diversity. Through extensive comparative experiments and analyses on two public datasets, we demonstrate the superior performance of our method.

2) To fully explore the joint distribution between audio information and gestures, we propose the Audio Feature Constructor (AFC). It employs a two-stage attention operation to extract features from the Mel-spectrogram, which are then combined with features from the raw audio signal. This approach enhances the model's capacity to learn and utilize relevant audio information.

3) To eliminate cumulative errors in non-autoregressive tasks, we introduce the Cross Dimension Transformer (CDformer). Additionally, we introduce a cross-dimension attention mechanism that focuses on the spatial and channel dimensions of input modalities, reducing the impact of redundant information on the model.

The rest of this article is structured as follows: Section 2 shows the work involved in co-speech gesture generation. In Section 3, we introduce the proposed Audio to Diffusion Gesture pipeline. Section 4 is the experimental part. Section 5 gives the conclusion.

2. Related work

Deep learning-based approaches for co-speech gesture generation primarily rely on three input modalities: Audio, text, and non-linguistic modalities [20]. We focused on audio-driven diffusion-based gesture generation. Therefore, in this section, we discuss audio-driven gesture generation and a diffusion-based motion synthesis mode.

2.1. Audio-driven co-speech gesture synthesis

Hasegawa et al. [21] proposed a set of audio-driven gesture generation methods based on bidirectional LSTM, incorporating time filtering to mitigate the discontinuity in the generated pose sequences. Kucherenko et al. [22] transformed audio input into 3D joint coordinates of gesture sequences while training a speech encoder to reduce the dimensionality of speech for motion representation. However, this approach overlooks the one-to-many relationship between audio and gestures. For example, a person might make different gestures for the same sentence at different times. Ginosar et al. [11] converted 2D spectrograms into 1D signals and employed generative adversarial networks to predict gestures. Ao et al. [23] introduce a co-speech gesture synthesis method using rhythm-based segmentation and hierarchical embeddings to align speech and gestures, achieving superior rhythmic and semantic coherence. Ye et al. [24] proposed an end-to-end flow-based model without style labels, combining a global encoder and gesture perceptual loss to generate natural gestures. Liu et al. [25] introduce BEAT, a large-scale motion capture dataset with semantic and emotional annotations, and propose a cascaded network (CaMN) for multi-modal gesture synthesis. Yi et al. [26] proposed a novel approach for generating realistic 3D body motions, hand gestures, and facial expressions directly from speech. The method leverages a new dataset and an innovative speech-

to-motion framework that independently models facial expressions and body-hand movements. Qian et al. [16] encoded Mel-spectrograms into template vectors to reduce uncertainty in synthesized poses. Our method conditions the generation of co-speech gestures on key features of both audio waveforms and Mel-spectrograms.

2.2. Diffusion-based motion synthesis mode

Recently, diffusion models [13,27,28] have made remarkable strides in generative modeling tasks, owing to techniques involving forward noise injection and reverse denoising. While these approaches may demand substantial computational resources, the generated samples demonstrate high quality and diversity. Zhang et al. [29] and Chen et al. [30] explored motion diffusion generation models conditioned on text. Alexanderson et al. [31] and Zhu et al. [32] were one of the first to synthesize co-speech gestures using the diffusion model. Yang et al. [33] proposed DiffuseStyleGesture, a diffusion-based model using attention mechanisms to generate high-quality, speech-matched, diverse, and stylized gestures. Yuan et al. [34] proposed a physics-based diffusion paradigm to guide motion generation. Ao et al. [35] proposed a neural network for stylized co-speech gesture synthesis using CLIP-guided multimodal prompts and a latent diffusion model, enabling flexible, realistic, and semantically aligned gesture generation. We employ a latent diffusion model to generate co-speech gestures, which alleviates the issue of random jitter in human motion synthesis tasks. We also adopt the concept of latent diffusion for generating co-speech gestures. However, unlike Ao et al. [35], we focus more on the key features within the audio data. In the gesture synthesis stage, the cross-dimension attention mechanism of CDformer is used to minimize the impact of redundant information on the model.

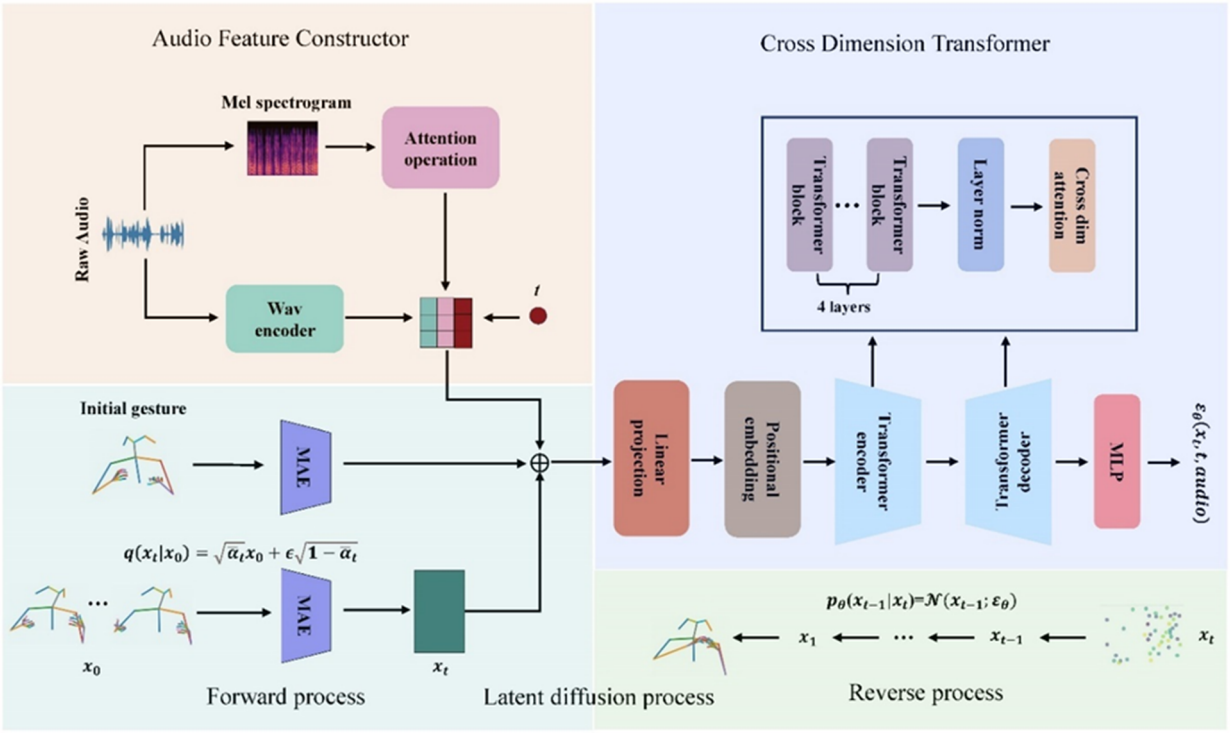


Figure 2. The proposed Audio to Diffusion Gesture(A2DG) generation pipeline.

3. Method

3.1. Preliminary

Our objective is to generate co-speech gestures that are more expressive and demonstrate higher fidelity. Given a N-frame co-speech video, pose sequences $x_0 = [s_1, \ldots, s_N]$ are extracted using pose estimators such as Openpose [36] and Expose [37]. To stabilize the training model, we aim at the skeleton deformation problem of different lengths. We follow the baseline method [8,10,32] and define the unit direction vector $s_i = [d_{i,1}, \ldots, d_{i,J-1}]$. Here, J is the total number of joints, and $d_{i,1}$ represents the direction vector between two skeletal key points of the J joint in 1th frame. The audio information matching the gesture is represented as $a = [a_1, \ldots, a_N]$ and is combined with time step t. The initial pose can facilitate a smoother synthesis process, we use the last 4 frames of the previously synthesized gestures as the seed gestures $M = [m_1, \ldots, m_4]$. We introduce a Motion Auto-Encoder (MAE) [8] that compresses both M and the noisy gesture x_0 into a lower-dimensional latent space, where the diffusion process produces latent data x_t. Finally, the reverse denoising is performed within our CDformer network to synthesize the gestures. The combination c of the above context information is input into our audio to a diffusion gesture generation framework G. Figure 2 illustrates the detailed process.

The end goal p can be described as:

$$p = G(c) \tag{1}$$

3.2. Diffusion model for gesture generation

Most research [8–11] is based on Generative Adversarial Networks [12] and applied to the task of audio-gesture, which is a complex mapping relation. Such training tends to be unstable and causes the mode to collapse. In order to generate high-quality and diverse gestures, we design an audio-driven gesture generation pipeline based on the diffusion model [13,28]. The core idea of the model is to train a probability model to eliminate the normal distribution noise step by step, which can be defined as $p_\theta(x_0) := \int p_\theta(x_{0:T})dx_{1:T}$, to approximate the real distribution $q(x_0)$, and then generate the target gesture, where $x_1 \sim x_T$ are the latent data.

The diffusion model is divided into two parts: forward diffusion process and reverse denoising.

Diffusion process: The forward diffusion process follows the Markov chain [27], and the model will gradually add Gaussian noise to the input data according to the variance schedule $\beta_t \in (0,1)$, until the input distribution approaches a posteriori distribution $\mathcal{N}(0, I)$:

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

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I). \tag{3}$$

where variance schedule $\beta_t \in (0,1)$ are hyper-parameters that follows the monotonically decreasing time table.

Denoising process: The noise gesture x_t is obtained by the erosion of the pose sequence X by the input noise in the diffusion process, and the denoising process follows the Markov chain. Reversing

the forward process $p_\theta(x_{0:T})$ allows sampling ground truth x_0 by starting from $p(x_T) = \mathcal{N}(x_T; 0, I)$, each step is a learning Gaussian transition $(\mu_\theta, \Sigma_\theta)$:

$$p_\theta(x_{0:T}) = p_\theta(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t), \tag{4}$$

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \tag{5}$$

During the training, we follow Ho et al. [9] to generate samples from more efficiently, which can be formulated as follows:

$$q(x_t \mid x_0) = \sqrt{\bar{\alpha}_t}x_0 + \epsilon\sqrt{1-\bar{\alpha}_t}, \epsilon \sim \mathcal{N}(0, I) \tag{6}$$

where $\alpha_t = 1 - \beta_t, \bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$. The above unconditional diffusion model can generate a better quality co-speech gesture, but additional conditions need to be injected to control the quality of the generated gesture. Therefore, it is necessary to construct a network to adapt to $\varepsilon_\theta(x_t, t, audio)$. Thus, we can use Eq (6) to generate the noise gesture x_t directly. At the time step of uniformly sampling the time point *t*, the audio feature vectors are extracted from the audio feature constructor, which contains more audio feature information. These conditions constitute context information *c*, which is input into our proposed CDformer for sequence modeling. We use Mean-Square-Error (MSE) loss to optimize the diffusion model parameters:

$$L(\theta) = E_{t\in[1,T],\mathbf{x}_0\sim q(\mathbf{x}_0),\varepsilon\sim\mathcal{N}(0,I)}[\| \varepsilon - \varepsilon_\theta(x_t, t, audio) \|_2^2] \tag{7}$$

3.3. Audio feature constructor

Mel-spectrograms, converted from audio signals, contain rich time-frequency information and align with the human auditory system's perception of audio. Therefore, we propose the AFC (Audio Feature Constructor), which enhances the model's capability to perceive and utilize global audio information by extracting audio features from both 1D audio signals and 2D Mel-spectrograms. Specifically, we employ the audio encoder from Yoon et al. [8], where the raw audio waveform is processed by cascaded 1D-CNNs to generate audio feature vectors. Because using vanilla 1D-CNN to process raw audio waveforms will limit our model to learning the joint distribution between audio and gestures, we concurrently apply a two-stage attention operation [38] to process the Mel-spectrogram. Given $\mathrm{M} \in \mathbb{R}^{\mathrm{N}\times\mathrm{C}\times\mathrm{F}\times\mathrm{T}}$ as input, which is converted from raw audio, C is the number of channels (set to 1 for mono audio), F is the frequency dimension, and T is the time dimension.

First stage: We introduce bilinear pooling [39] to capture global audio features in the Mel-spectrogram. Bilinear pooling performs summation pooling on all pairs of audio feature vectors (x_i, y_i) in the Mel-spectrogram to extract key audio features:

$$G_{\text{bilinear}}(X, Y) = XY^\top = \sum_{\forall i} x_i y_i^\top \tag{8}$$

where X and Y are audio feature maps from the same time domain. We use a softmax attention map to collect key audio features from different locations into a set of global descriptors.

Second stage: As shown in Figure 3, for each time-frequency input position $i = 1, \ldots, FT$, an attention vector is generated based on the local audio feature v_i. This attention vector supplements the global audio feature information with key audio features from the global descriptors, resulting in the final audio feature vector z_i.

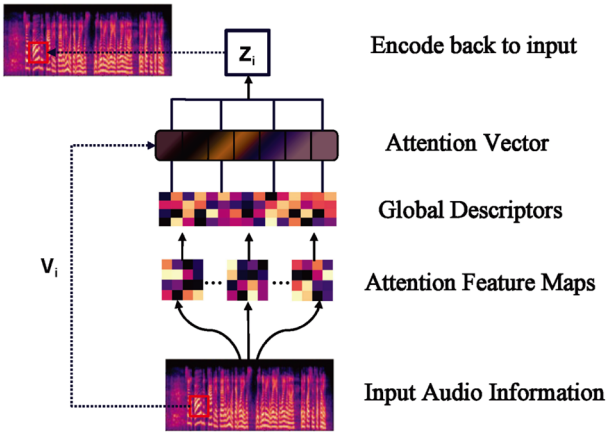


Figure 3. Two stage audio attention operation for Mel-spectrogram.

3.4. Cross dimension transformer

In this section, we combine the initial gestures, time steps, and useful audio features to form the contextual information, which is concatenated with the noise gesture sequence along the feature channels to create condition tokens. These tokens are then fed into our proposed denoising model, CDformer (Cross Dimension Transformer). As illustrated in Figure 2, after linear projection, the input embedding dimension is adjusted to the hidden layer dimension:

$$y = Wx + b \tag{9}$$

where W is the weight matrix, b is the bias vector, x is the input vector, and y is the output vector. The positional embedding parameters provide a unique embedding vector for each gesture, enabling the model to capture the positional information of gestures. We apply the Vision Transformer [40] encoder-decoder network to denoise the noisy gestures. The conditional tokens pass through hierarchical transformer blocks, where the multi-head mechanism splits the input tokens into multiple parts and processes them using the self-attention mechanism:

$$\text{Attention}\ (Q, K, V) = \sigma\left(\frac{Q \cdot K^T}{\sqrt{d}}\right)V \tag{10}$$

where σ is the softmax operator, Q is the query feature vector, K is the key feature vector, V is the value feature vector, and d is the channel dimension of the gesture features.

When concatenating audio information along the feature channels, redundant information may arise due to the overlap between the two audio modalities. Therefore, we introduce cross-dimension attention [41] to mitigate the impact of this redundancy on the model. By focusing on both spatial and

channel information, this mechanism enables the model to autonomously select the most important features for learning. This operation is applied after the layer normalization (layer norm) in the encoder-decoder architecture:

$$\text{CD-Attn(token)}=\frac{1}{3}([V_a \oplus V_m \oplus P_t]) \tag{11}$$

where V_a represents features extracted from raw audio, V_m represents features extracted from the Mel-spectrogram, P_t denotes time embedding features, and $\oplus$ denotes the concatenation operation.

4. Experiments and details

4.1. Co-speech gesture dataset

We refrained from utilizing co-speech gestures collected in a studio environment, as requiring speakers to produce gestures that perfectly align with their speech often results in exaggerated and insincere expressions [10]. Such an approach contradicts our research objective, which is to acquire naturally fluent and rhythmical co-speech gestures.

TED Gesture. TED Gesture is a large-scale dataset for co-speech gesture generation, featuring 1776 TED talk videos with various narrators and topics. The dataset includes the 3D poses of speaker's upper bodies and the corresponding audio sequences. Following the data processing approach of previous works [8,10,32], we resampled human poses at 15 FPS (approximately 4 seconds per sample). Each video sequence is 34 frames long with a step length of 10 frames. The upper body posture includes 10 key points, resulting in a total of 252,109 training samples. These samples are divided into training, validation, and test sets in an 80, 10, and 10% split, respectively.

TED-Expressive. High-quality finger motion data is essential for generating expressive and meaningful gestures [20], yet it is rare in existing datasets. Building on TED Gesture, TED-Expressive [10] annotates 43 key points on the speaker's upper body, including 13 upper body joints and 30 finger joints, using the 3D pose estimator ExPose [37]. The other settings are consistent with TED Gesture.

4.2. Evaluation metrics

To objectively evaluate the proposed pipeline, we use three common objective evaluation indicators to measure the quality of co-speech gesture generation.

Fréchet Gesture Distance (FGD). FGD refers to the distribution distance between synthesized gestures and ground truth in the latent feature space. The closer the distribution distance, the more similar the synthetic gesture is to the real one, which is similar to the Fréchet Inception Distance (FID) [42] definition in image generation studies and is the main index to evaluate the rationality of gesture generation. Yoon et al. [8] trained a feature extractor on the Human3.6M [43] dataset for calculating the potential feature X of a real gesture and the potential feature $\hat{X}$ of a synthetic gesture:

$$\text{FGD}(X,\hat{X}) = \| \mu_r - \mu_g \|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}) \tag{12}$$

where μ_r and Σ_r are the first and second moments of the latent feature distribution Z_r of real human

gestures X, while μ_r and Σ_r come from the generated gestures, and μ_g and Σ_g are the first and second moment of the latent feature distribution Z_g of generated gestures $\hat{X}$. Note that to be fair to the trial on the Ted Gesture and TED-Expressive datasets, we did not train the feature extractor and used the one provided by Yoon et al. [8] and Liu et al. [10].

Beat Consistency Score (BC). To measure the correlation between synthetic gesture sequence and audio, Li et al. [44] proposed Beat Consistency Score (BC). Because of the differences in the kinematic velocities of human joints, it is necessary to calculate the mean absolute angle change (MAAC) of the angle θ_j between adjacent frames by:

$$\mathrm{MAAC}(\theta_j) = \frac{\sum_{s=1}^{S}\sum_{t=1}^{T-1} \| \theta_{j,s,t+1} - \theta_{j,s,t} \|_1}{S * (T-1)} \tag{13}$$

where S represents the total number of clips in the dataset and T represents the number of frames in each clip.

We follow Li et al. [44] to calculate the kinematic beat as the local minimum of the kinemat velocity. BC computes the average distance between every audio beat and its nearest kinematic beat with the following equation:

$$BC = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-\frac{\min_{\forall b_j^x \in B^x} \| b_i^x - b_j^y \|^2}{2\sigma^2} \right) \tag{14}$$

where $B^x = \{b_i^x\}$ is the kinematic beats, $B^y = \{b_j^y\}$ is the audio beat, and σ is a parameter to normalize sequences: $\sigma = 0.1$ empirically.

Diversity Score. Diversity is used to assess the degree of variation between the generating motions corresponding to the input [45]. We use the pre-trained autoencoder to capture the potential features of the synthesized gestures and calculate the average distances, randomly select 500 synthesized gestures, and calculate the average absolute error between the feature and the random feature.

4.3. Implementation details

Experimental environment. The software and hardware environment used in this experiment is shown in Table 1.

Table 1. Experimental environment configuration.

Hardware/Software	Configuration description
Operating System	Ubuntu 18.04 LTS
DeepLearning Framework	Pytorch 1.13.0
Programming Language	Python 3.7
CUDA Version	11.7
Processor	Intel Core i5-13600K
GPU	NVIDIA GeForce RTX 4090

Baselines. We select six state-of-the-art models in recent years to compare with the A2DG proposed in this paper. 1) Seq2Seq [17] follows the Encoder-Decoder structure to generate co-gesture from speech text; 2) speech2Gesture [11] converts the 2D spectrum of an audio signal into a 1D signal as input to generate a co-speech gesture; 3) Joint Embedding [18] maps text and motion into the same Embedding space, which is a representative work of text-generating motion; 4) Trimodal [8] uses audio, text, and speaker identity as context input, and introduces adversarial scheme training method to generate co-speech gesture; 5) HA2G [10] proposes a hierarchical audio-gesture generator across multiple level semantic granularity; and 6) DiffGesture [32] generates co-speech gesture using diffusion gesture stabilizer and annealed noise sampling strategy. These methods were evaluated by training on the TED Gesture dataset and TED-Expressive dataset.

Experimental details. For a fair comparison, we followed the previous work [8,10,32], setting N = 34 and M = 4, where N indicates the sequence in which the sample split into 34 frames and M indicates the first four frames as seed postures. For subdivision strides S = 10, J = 10 for upper-body joint training on the TED Geture dataset, and J = 43 for upper-body joint (especially for finger joints) training on the TED Expressive dataset, the position of the joint is represented by the normalized unit vector as the direction vector. For the diffusion model, we apply the denoising step T = 500, and the variance schedule is linearly increasing from 0.0001 to 0.02. The Cross Dimensional Transformer consists of a 4-layers transformer encoder with self-attention and feed-forward network and a 4-layer transformer decoder with a similar structure. The hidden dimension of the transformer blocks is set to 256 for TED Gesture and 512 for TED Expressive. We used the Adam Optimizer for model optimization, where $\beta_1 = 0.5$, $\beta_2 = 0.999$, and the learning rate is set to 0.0005 and 0.0002 for TED Gesture and TED Expressive, respectively. The model was trained on a single Nvidia GeForce RTX 4090 GPU, batch size = 128, and it takes about 9 hours for TED Gesture and about 14 hours for TED Expressive.

4.4. Results and analysis

Objective results and comparison. The quantitative results are shown in Table 2. We compare our approach with prior works on two publicly co-speech gesture datasets. As can be seen, our method achieves state-of-the-art performance on both FGD and Diversity metrics, surpassing the previous best approaches. FGD shows an improvement of 16.8% on TED Gesture, while a significant increase of 56% on TED Expressive, Diversity scores also increased by 1.406 and 0.376, respectively, indicating that our method can generate diverse and high-fidelity gestures. The BC scores we obtained for TED Gesture are lower compared to DiffGesture. It is noteworthy that BC and Diversity are meaningful only when synthesizing smooth and natural motions. However, synthesized gestures may exhibit irregular random jitter, leading to BC and Diversity scores surpassing Ground truth. For instance, in Table 2, the BC score for TED Expressive Ground truth is **0.703**, and the Diversity score is **178.827**, while DiffGesture scores higher in both metrics with **0.718** and **182.757**, respectively, surpassing Ground truth. This could be related to the abundant human joint points in the TED Expressive dataset. Moreover, the Fréchet Gesture Distance exhibits a high degree of statistical correlation with human similarity ratings from large-scale user studies [40]. Hence, apart from Fréchet Gesture Distance, other quantitative metrics should be used as references because they do not always align with the human perception of visual quality [31,46].

Table 2. The quantitative results comparison for TED Gesture and TED Expressive. ↓ denotes the lower the better, and ↑ denotes the higher the better. The best results are in bold.

	TED Gesture [8,17]			TED Expressive [10]		
Methods	FGD ↓	BC ↑	Diversity ↑	FGD ↓	BC ↑	Diversity ↑
Ground truth	0	0.698	108.525	0	0.703	178.827
Attention Seq2Seq [17]	18.154	0.196	82.776	54.920	0.152	122.693
Speech2Gesture [11]	19.254	0.668	93.802	54.650	0.679	142.489
Joint Embedding [18]	22.083	0.200	90.138	64.555	0.130	120.627
Trimodal [8]	3.729	0.667	101.247	12.613	0.563	154.088
HA2G [10]	3.072	0.672	104.322	5.306	0.641	173.899
DiffGesture [32]	1.506	**0.699**	106.722	2.600	**0.718**	182.757
A2DG (Ours)	**1.253**	0.678	**108.128**	**1.126**	**0.718**	**183.133**

Subjective results and comparison. The gesture visualization result is illustrated in Figure 4. We chose to contrast it with DiffGesture [32], which performed the best among numerous baselines. Ground truth exhibits less variation, lacking diversity or rhythmic gestures. While DiffGesture initially displays good continuity in synthetic gestures, its later transitions into monotonous gestures, with a slight swing of the right arm marked by a red oval and a consistently drooping left arm showing irregular shaking marked by an orange oval. This aligns with our quantitative analysis findings. Our synthesized gestures avoid rigid movement patterns, with relatively smooth transitions between them. Moreover, when descriptive terms like "something negative happens" are present, our gestures demonstrate a level of semantic relevance.

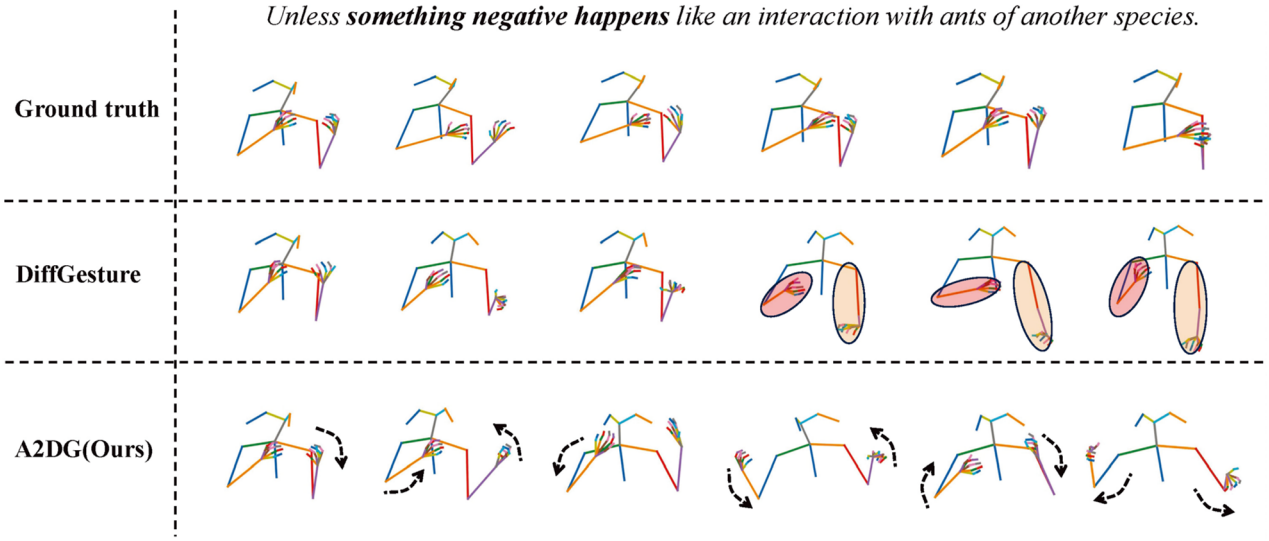
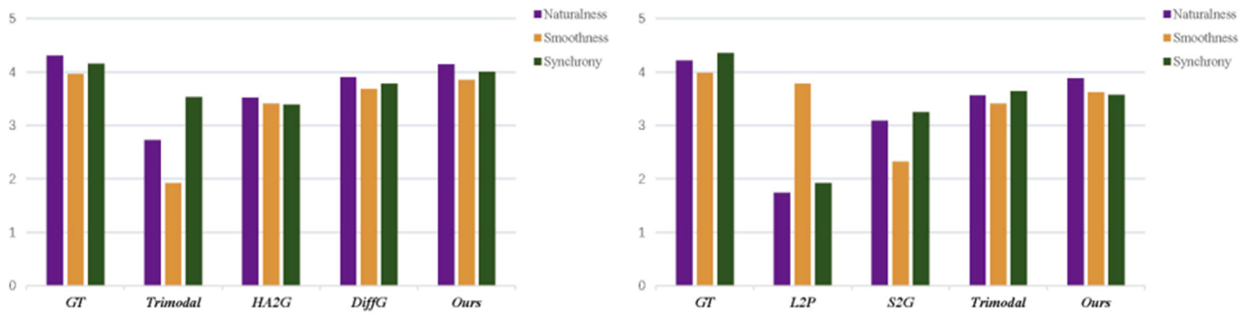


Figure 4. The visualization subjective results of synthesized gesture sequence.

User Study. The performance of the generative model cannot be accurately measured by objective metrics alone [18]. For instance, in the same contextual scenario, as shown in Figure 4, DiffGesture presents a segment of dull gestures in the later stages, whereas A2DG (ours) generates more expressive gestures. Hence, it is imperative to integrate human judgment and objective metrics in assessing gesture generation models. We conducted a user study that compared our proposed pipeline with several baselines [8,10,11,18,32], assessment is conducted based on three aspects of gestures: Naturalness, Smoothness and Synchrony. Specifically, we generated gesture sequences from the TED Gesture and Ted Expressive test dataset and randomly selected 10 slices of approximately 20 seconds. We asked 10 participants to rate the slices after watching them twice. Scores range from 1 to 5, where higher scores indicate greater participant endorsement of gestures synthesized by the model. As shown in Figure 5(a), our approach performs well for all three metrics, for TED Expressive, gesture generation approaches Ground Truth levels in terms of Naturalness, Smoothness, and Synchrony. This may be attributed to the dataset's richness in finger-joint information, resulting in higher-quality synthesized gestures. For TED Gesture, with only 10 upper body joints, gesture synthesis quality is comparatively lower, as depicted in Figure 5(b). Our approach excels other baselines in Naturalness but falls short of L2P in Smoothness and trails behind Trimodal in Synchrony.



(a) User study results on the TED-Expressive dataset

(b) User study results on the TED-Gesture dataset

Figure 5. The statistical results of our user study on TED-Expressive dataset and TED-Gesture dataset. On a scale of 1–5, the higher the better.

4.5. Ablation studies

Quantitative ablation study. To validate the effectiveness of the proposed components in our method, we conduct ablation studies on TED Gesture and TED Expressive datasets, and the quantitative ablation results are shown in Table 3. The removal of AFC leads to varying degrees of decline across three evaluation metrics on both datasets, demonstrating the noticeable improvement in our model's ability to learn the joint distribution of audio and gestures after extracting useful audio information through AFC. After removing CDformer, apart from the improvement in the Diversity score on TED Expressive, the remaining two metrics degrade. Mainly benefiting from the robust sequence modeling capability of CDformer, it can impact the quality of generated gestures. Furthermore, given that the FGD metric currently best aligns with human perception among all objective evaluation measures, the significant decrease in FGD after removing our proposed components indicates the effectiveness of our approach in synthesizing high-quality co-speech gestures.

Table 3. The results of quantitative ablation study regarding the proposed modules.

	TED Gesture [8,17]			TED Expressive [10]		
Methods	FGD ↓	BC ↑	Diversity ↑	FGD ↓	BC ↑	Diversity ↑
Ground truth	0	0.698	108.525	0	0.703	178.827
w/o AFC	1.803	0.662	105.389	1.339	0.713	177.642
w/o CDformer	1.325	0.661	107.220	1.453	0.717	180.760
A2DG (Ours)	**1.253**	**0.678**	**108.128**	**1.126**	**0.718**	**183.133**

Qualitative ablation study. We conduct a qualitative ablation study on the proposed modules, the results are shown in Figure 6. Without our Audio Feature Constructor, simply injecting raw audio information into the network degrades the quality of the synthesized gestures. We highlight the unnatural gestures generated by our network within the red box. Our complete pipeline synthesizes diverse and meaningful gestures. For instance, when saying "now depend", our synthesized gesture extends the arm to emphasize the stressed word.

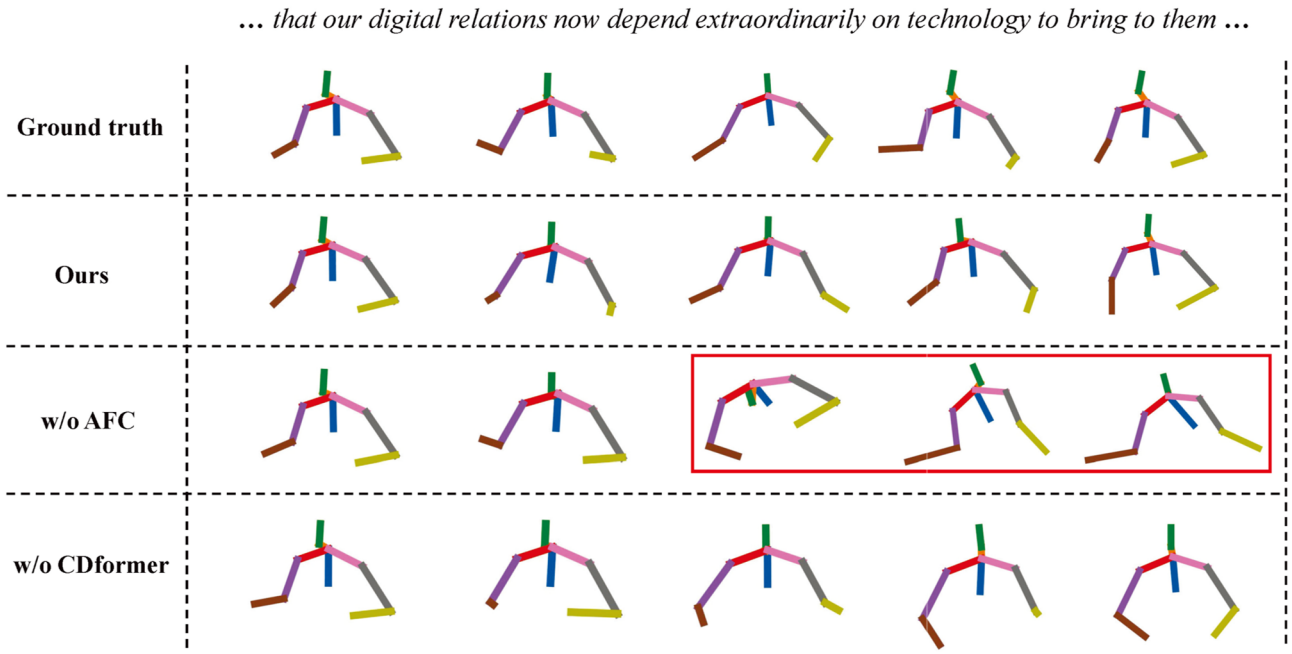


Figure 6. The visualization qualitative ablation results of synthesized gesture sequence.

5. Conclusions

In this paper, we propose a diffusion model-based audio-driven co-speech gesture generation framework comprising two modules: AFC and CDformer. The AFC module extracts useful audio feature information from raw audio waveforms and Mel-spectrograms, enhancing the model's ability to learn the joint distribution between audio and gestures. The cross-dimension attention in the CDformer module focuses on spatial and channel information, thereby reducing the impact of redundant information on the model. Leveraging the powerful sequence modeling capabilities of the

Transformer, our method can generate diverse and realistic gestures.

Our research is limited to upper body movements, and during the synthesis phase, speaker identity was not incorporated to generate gestures with personal style. Therefore, in future work, we plan to explore additional audio cues, such as analyzing prosodic features and extracting speaker specific characteristics from Mel-spectrograms, to generate stylized full-body co-speech gestures.

Use of AI tools declaration

The authors declare they have not used artificial intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions. This work is supported by the Fujian Province Industrial Guidance (Key) Project (Grant No. 2022H0053), the Sanming Major Science and Technology Project of Industry-University-Research Collaborative Innovation (Grant No. 2022-G-4) and the Start-up Research Project of Fujian University of Technology (Grant Nos. GY-Z21064, GY-Z21065).

Conflict of interest

The authors declare there are no conflicts of interest.

References

1. S. Van Mulken, E. André, J. Müller, The Persona Effect: How Substantial Is It?, in *People and Computers XIII : Proceedings of HCI'98*, Springer London, (1998), 53–66. https://doi.org/10.1007/978-1-4471-3605-7_4
2. J. Cassell, D. McNeill, K. E. McCullough, Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information, *Pragmatics Cognit.*, **7** (1999), 1–34. https://doi.org/10.1075/pc.7.1.03cas
3. T. Kucherenko, P. Jonell, Y. Yoon, P. Wolfert, A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA challenge 2020, in *26th International Conference on Intelligent User Interfaces (IUI)*, (2021), 11–21. https://doi.org/10.1145/3397481.3450692
4. C. M. Huang, B. Mutlu, Robot behavior toolkit: generating effective social behaviors for robots, in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, (2012), 25–32. https://doi.org/10.1145/2157689.2157694
5. M. Salem, S. Kopp, I. Wachsmuth, K. Rohlfing, F. Joublin, Generation and evaluation of communicative robot gesture, *Int. J. Social Rob.*, **4** (2012), 201–217. https://doi.org/10.1007/s12369-011-0124-9
6. A. Kranstedt, S. Kopp, I. Wachsmuth, Murml: A multimodal utterance representation markup language for conversational agents, in *AAMAS'02 Workshop Embodied Conversational Agents-Let's Specify and Evaluate Them!*, 2002.

7. J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, et al., 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 (SIGGRAPH)*, (1994), 413–420. https://doi.org/10.1145/192161.192272
8. Y. Yoon, B. Cha, J. H. Lee, M. Jang, J. Lee, J. Kim, et al., Speech gesture generation from the trimodal context of text, audio, and speaker identity, *ACM Trans. Graphics*, **39** (2020), 1–16. https://doi.org/10.1145/3414685.3417838
9. U. Bhattacharya, E. Childs, N. Rewkowski, D. Manocha, Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning, in *Proceedings of the 29th ACM International Conference on Multimedia (MM)*, (2021), 2027–2036. https://doi.org/10.1145/3474085.3475223
10. X. Liu, Q. Wu, H. Zhou, Y. Xu, R. Qian, X. Lin, et al., Learning hierarchical cross-modal association for co-speech gesture generation, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 10452–10462. https://doi.org/10.1109/CVPR52688.2022.01021
11. S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, J. Malik, Learning individual styles of conversational gesture, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 3492–3501. https://doi.org/10.1109/CVPR.2019.00361
12. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., Generative adversarial nets, *Commun. ACM*, **63** (2020), 139–144. https://doi.org/10.1145/3422622
13. J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS)*, (2020), 6840–6851.
14. S. Alexanderson, G. E. Henter, T. Kucherenko, J. Beskow, Style-controllable speech-driven gesture synthesis using normalising flows, *Comput. Graphics Forum*, **39** (2020), 487–496. https://doi.org/10.1111/cgf.13946
15. T. Kucherenko, P. Jonell, S. Van Waveren, G. E. Henter, S. Alexandersson, I. Leite, et al. Gesticulator: A framework for semantically-aware speech-driven gesture generation, in *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI)*, (2020), 242–250. https://doi.org/10.1145/3382507.3418815
16. S. Qian, Z. Tu, Y. Zhi, W. Liu, S. Gao, Speech drives templates: Co-speech gesture synthesis with learned templates, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 11057–11066. https://doi.org/10.1109/ICCV48922.2021.01089
17. Y. Yoon, W. R. Ko, M. Jang, J. Lee, J. Kim, G. Lee, Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots, in *2019 International Conference on Robotics and Automation (ICRA)*, (2019), 4303–4309. https://doi.org/10.1109/ICRA.2019.8793720
18. C. Ahuja, L. P. Morency, Language2pose: Natural language grounded pose forecasting, in *2019 International Conference on 3D Vision (3DV)*, (2019), 719–728. https://doi.org/10.1109/3DV.2019.00084
19. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, (2017), 6000–6010. https://doi.org/10.48550/arxiv.1706.03762
20. S. Nyatsanga, T. Kucherenko, C. Ahuja, G. E. Henter, M. Neff, A comprehensive review of data-driven co-speech gesture generation, *Comput. Graphics Forum*, **42** **(**2023), 569–596. https://doi.org/10.1111/cgf.14776

21. D. Hasegawa, N. Kaneko, S. Shirakawa, H. Sakuta, K. Sumi, Evaluation of speech-to-gesture generation using Bi-directional LSTM network, in *Proceedings of the 18th International Conference on Intelligent Virtual Agents (IVA)*, (2018), 79–86. https://doi.org/10.1145/3267851.3267878
22. T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, H. Kjellström, Analyzing input and output representations for speech-driven gesture generation, in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents (IVA)*, (2019), 97–104. https://doi.org/10.1145/3308532.3329472
23. T. Ao, Q. Gao, Y. Lou, B. Chen, L. Liu, Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings, *ACM Trans. Graphics*, **41** (2022), 1–19. https://doi.org/10.1145/3550454.3555435.
24. S. Ye, Y. H. Wen, Y. Sun, Y. He, Z. Zhang, Y. Wang, et al., Audio-driven stylized gesture generation with flow-based model, in *European Conference on Computer Vision*, **13665** (2022), 712–728. https://doi.org/10.1007/978-3-031-20065-6_41
25. H. Liu, Z. Zhu, N. Iwamoto, Y. Peng, Z. Li, Y. Zhou, et al., BEAT: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis, in *European Conference on Computer Vision*, **13667** (2022), 612–630. https://doi.org/10.1007/978-3-031-20071-7_36
26. H. Yi, H. Liang, Y. Liu, Q. Cao, Y. Wen, T. Bolkart, et al., Generating holistic 3D human motion from speech, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2023), 469–480.
27. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 10684–10695. https://doi.org/10.48550/arxiv.2112.10752
28. P. Dhariwal, A. Nichol, Diffusion models beat GANs on image synthesis, in *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS)*, (2021), 8780–8794. https://doi.org/10.48550/arXiv.2105.05233
29. M. Zhang, Z. Cai, L. Pan, F. Hong, X, Guo, L, Yang, et al., MotionDiffuse: Text-driven human motion generation with diffusion model, *IEEE Trans. Pattern Anal. Mach. Intell.*, **46** (2024), 4115–4128. https://10.1109/TPAMI. 2024.3355414
30. X. Chen, B. Jiang, W. Liu, Z. Huang, B. Fu, T. Chen, et al., Executing your commands via motion diffusion in latent space, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2023), 18000–18010. https://10.1109/CVPR52729.2023.01726
31. S. Alexanderson, R. Nagy, J. Beskow, G. E. Henter, listen, denoise, action! audio-driven motion synthesis with diffusion models, *ACM Trans. Graphics*, **42** (2023). https://doi.org/10.1145/3592458
32. L. Zhu, X. Liu, X. Liu, R. Qian, Z. Liu, L. Yu, Taming diffusion models for audio-driven co-speech gesture generation, in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2023), 10544–10553. https://doi.org/10.1109/CVPR52729.2023.01016
33. S. Yang, Z. Wu, M. Li, Z. Zhang, L. Hao, W. Bao, et al., DiffuseStyleGesture: stylized audio-driven co-speech gesture generation with diffusion models, in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, **650** (2023), 5860–5868. https://doi.org/10.24963/ijcai.2023/650

34. Y. Yuan, J. Song, U. Iqbal, A. Vahdat, J. Kautz, PhysDiff: Physics-guided human motion diffusion model, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2023), 16010–16021. https://doi.org/10.48550/arxiv.2212.02500
35. T. Ao, Z. Zhang, L. Liu, GestureDiffuCLIP: Gesture diffusion model with CLIP latents, *ACM Trans. Graphics*, **42** (2023), 1–18. https://doi.org/10.1145/3550454.3555435
36. Z. Cao, T. Simon, S. Wei, Y. Sheikh, Realtime multi-person 2D pose estimation using part affinity fields, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 1302–1310. https://doi.org/10.1109/CVPR.2017.143
37. V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, M. J. Black, Monocular expressive body regression through body-driven attention, in *16th European Conference Computer Vision (ECCV)*, **12355** (2020), 20–40. https://doi.org/10.1007/978-3-030-58607-2_2
38. Y. Chen, Y. Kalantidis, J. Li, S. Yan, J. Feng, A^2-Nets: Double attention networks, in *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS)*, (2018), 350–359. https://doi.org/10.48550/arxiv.1810.11579
39. T. Y. Lin, A. RoyChowdhury, S. Maji, Bilinear CNN models for fine-grained visual recognition, in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, (2015), 1449–1457. https://doi.org/10.1109/ICCV.2015.170
40. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is worth 16×16 words: Transformers for image recognition at scale, preprint, arXiv:2010.11929v2. https://doi.org/10.48550/arXiv.2010.11929
41. D. Misra, T. Nalamada, A. U. Arasanipalai, Q. Hou, Rotate to attend: Convolutional triplet attention module, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, (2021), 3139–3148. https://doi.org/10.1109/WACV48630.2021.00318
42. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs trained by a two time-scale update rule converge to a local nash equilibrium, in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, (2017), 6629–6640. https://doi.org/10.48550/arXiv.1706.08500.
43. C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments, *IEEE Trans. Pattern Anal. Mach. Intell.*, **36** (2013), 1325–1339. https://doi.org/10.1109/TPAMI.2013.248
44. R. Li, S. Yang, D. A. Ross, A. Kanazawa, AI choreographer: Music conditioned 3D dance generation with AIST++, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 13401–13412. https://doi.org/10.1109/ICCV48922.2021.01315
45. H. Y. Lee, X. Yang, M. Y. Liu, T. C. Wang, Y. D. Lu, M. H. Yang, et al., Dancing to music, in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, (2019), 3586–3596. https://doi.org/10.48550/arxiv.1911.02001
46. T. Kucherenko, P. Wolfert, Y. Yoon, C. Viegas, T. Nikolov, M. Tsakov, et al., Evaluating gesture generation in a large-scale open challenge: The GENEA Challenge 2022, *ACM Trans. Graphics*, **43** (2024). https://doi.org/10.1145/3656374