Co-Speech Gesture Video Generation via Motion-Decoupled Diffusion Model

Xu He[1] Qiaochu Huang[1] Zhensong Zhang[2] Zhiwei Lin[1] Zhiyong Wu[✉,1,4]
Sicheng Yang[1] Minglei Li[3] Zhiyi Chen[3] Songcen Xu[2] Xiaofei Wu[2]
[1] Shenzhen International Graduate School, Tsinghua University [2] Huawei Noah's Ark Lab
[3] Huawei Cloud Computing Technologies Co., Ltd [4] The Chinese University of Hong Kong
{hex22, hqc22, lzw22, yangsc21}@mails.tsinghua.edu.cn zywu@sz.tsinghua.edu.cn
{zhangzhensong, liminglei29, chenzhiyi2, xusongcen, wuxiaofei2}@huawei.com

Figure 1. **Examples of our generated gesture videos.** White dashed arrows indicate gestures corresponding to bold words.

Abstract

Co-speech gestures, if presented in the lively form of videos, can achieve superior visual effects in human-machine interaction. While previous works mostly generate structural human skeletons, resulting in the omission of appearance information, we focus on the direct generation of audio-driven co-speech gesture videos in this work. There are two main challenges: 1) A suitable motion feature is needed to describe complex human movements with crucial appearance information. 2) Gestures and speech exhibit inherent dependencies and should be temporally aligned even of arbitrary length. To solve these problems, we present a novel motion-decoupled framework to generate co-speech gesture videos. Specifically, we first introduce a well-designed nonlinear TPS transformation to obtain latent motion features preserving essential appearance information. Then a transformer-based diffusion model is proposed to learn the temporal correlation between gestures and speech, and performs generation in the latent motion space, followed by an optimal motion selection module to produce long-term coherent and consistent gesture videos. For better visual perception, we further design a refinement network focusing on missing details of certain areas. Extensive experimental results show that our proposed framework significantly outperforms existing approaches in both motion and video-related evaluations. Our code, demos, and more resources are available at https://github.com/thuhcsi/S2G-MDDiffusion.

1. Introduction

Co-speech gestures, as a typical form of non-verbal behavior [7], convey a wealth of information and play an important role in human communication. Appropriate gestures complement human speech and thus benefit comprehension, persuasion, and credibility [46]. Hence providing artificial agents with human-like and speech-appropriate gestures is crucial in human-machine interaction.

To achieve this goal, several methods have been developed for automatic co-speech gesture generation, with a particular focus on deep learning techniques. However, they mostly aim at generating gestures as 2D/3D human skeletons. While relatively easy to generate, skeletons totally discard appearance information and create a disparity with human perception [23]. As a result, they need to be further processed for better visualization. For example, some work binds skeletons to custom virtual avatars and manually renders them using software like Blender and Maya, consuming exhaustive human labor. Other studies [13, 27] train independent image synthesizers [4] to translate skeletons into animated images, which still rely on hand-crafted

annotations and yield noticeable inter-frame jitters.

Different from previous methods that only generate skeletons, we aim to generate audio-driven co-speech gesture videos directly in a unified framework, which is challenging due to the following two reasons: First, we need to find a suitable motion feature that can describe both intricate motion trajectories and complex human appearance. A straightforward way is to design a two-stage pipeline by first generating hand-crafted and pre-defined skeletons as motion features and then synthesizing animated images with them. However, skeletons only contain positions of sparse joints and will lead to texture loss and accumulated errors, making it unsuitable for our task. Another way is to customize popular conditional video generation methods [11, 18, 30, 44] to solve our problem. These methods usually encode videos into a latent space and then generate content within this space using UNet-based diffusion models [14, 15, 33, 47]. However, they primarily concentrate on general video generation with latent features derived from VAEs lacking well-defined meaning and struggling to filter and retain necessary video information effectively. Directly applying them to videos concerning human motion results in implausible movements and missing fine-grained parts [30]. Second, gesture videos should be temporally aligned with the input audio even of arbitrary length, while it is still difficult to capture the inherent temporal dependencies between gestures and speech. Besides, existing video generation methods [30, 56] can only generate videos of fixed length, for example, 2 seconds. Generating longer consistent videos is either time-consuming or even impossible, since it requires much more computational resources.

To address these challenges, in this paper, we propose a novel unified motion-decoupled framework for audio-driven co-speech gesture video generation. The overview of our method is shown in Fig. 2. To decouple motion from gesture videos while preserving critical appearance information of body regions, we first carefully design a thin-plate spline (TPS) [5, 55] transformation to model first-order motion, which is nonlinear and thus flexible enough to adapt to curved human body regions. To be specific, we predict several groups of keypoints to generate TPS transformations, subsequently employed for estimating optical flow and guiding image warping to generate corresponding gesture video frames. Note that, gathered keypoints are considered as latent motion features, which allow for the explicit modeling of motion while maintaining a small scale, easing the burden on the generation model. Then we introduce a transformer-based diffusion model for generation within the latent motion space, equipped with self-attention and cross-attention modules to better capture the temporal dependency between speech and motion. To further extend the duration of generated videos, we propose an optimal motion selection module, which considers both coherence and consistency and helps to produce long-term gesture videos. Finally, for better visual quality, we present a UNet-like [29] refinement network supplemented with residual blocks [52] to capture local and global information of video frames, drawing more attention to certain regions and recovering missing details of appearance and textures.

To summarize, the main contributions of our works are as follows:

- We present a novel motion-decoupled framework to directly generate co-speech gesture videos in an end-to-end manner independent of hand-crafted structural human priors, where a nonlinear TPS transformation is used to extract latent motion features and ultimately guide the synthesis of gesture video frames.
- We design a transformer-based diffusion model on latent motion features, capturing temporal correlation between speech and gestures, which is followed by an optimal motion selection module concerning coherence and consistency. With both modules, we can generate diverse long co-speech gesture videos.
- We introduce a refinement network to allocate additional attention to certain areas and enhancing appearance and texture details, which is crucial for human perception.
- Extensive experimental results show that our framework can generate vivid, realistic, speech-matched, and long-term stable gesture videos of high quality that significantly outperform existing methods.

2. Related Works

Gesture generation on human skeletons. Early works consider gesture generation as an end-to-end regression task [2, 24] and tend to generate averaged gestures without diversity. Subsequent insights into the many-to-many relationship between speech and gestures prompt the adoption of diverse generation methods including GANs [10], VAEs [20], and Flows [3]. More currently, diffusion models excel at modeling complex data distribution and have emerged as a promising approach to generate gestures [39, 48, 54]. However, all these works depend on annotated datasets to generate human skeletons, including datasets labeled by pose estimators [1, 2, 50] and MoCap datasets [12, 22], suffering from error accumulation or insufficient data and totally devoid of appearance information. On the contrary, our framework generates gestures directly in the video form without relying on annotated skeleton priors.

Gesture video generation. To date, only a few works have made initial explorations into the problem of generating gesture videos directly. Zhou *et al*. [57] convert gesture video generation into a reenactment task and complete it in a rule-based way. They establish a motion graph with a reference video and search for a path matching the speech based on audio onset and a predefined keyword dictionary. However, it fails to generate novel gestures, and crafting

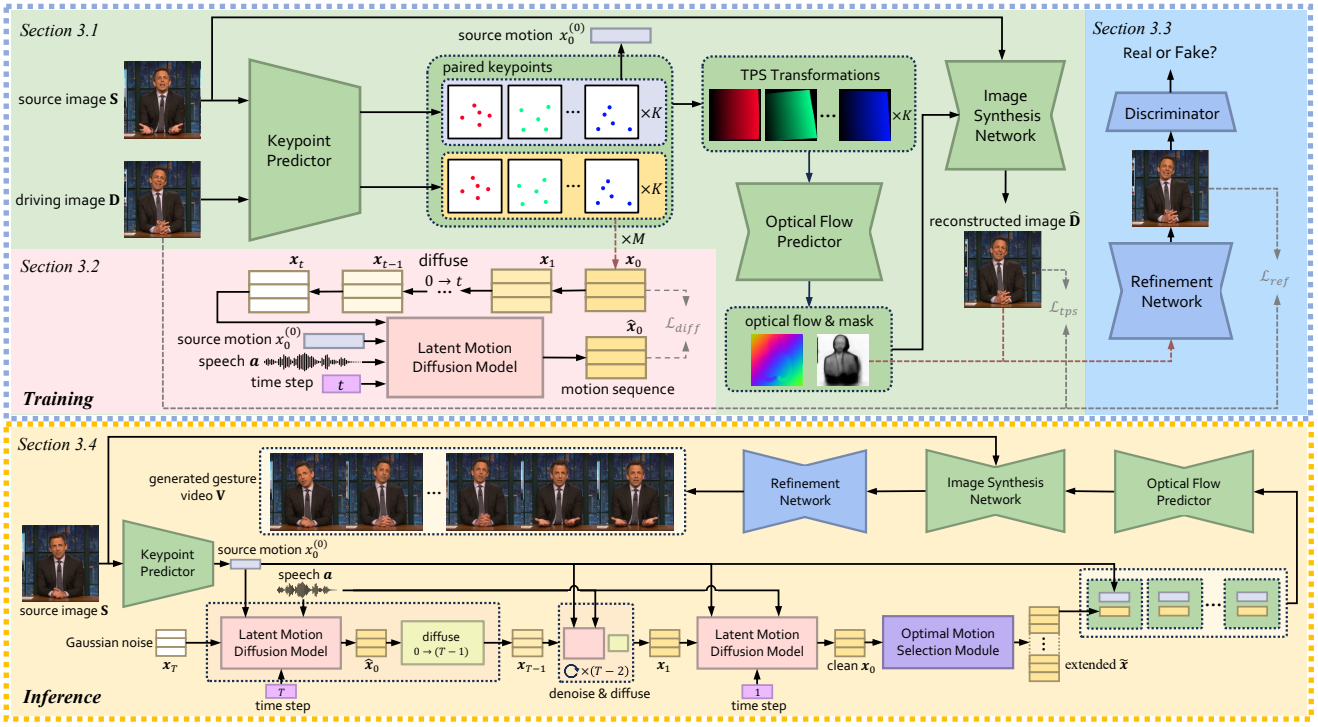

Figure 2. **Gesture video generation pipeline of our proposed framework** is composed of three core components: 1) the motion decoupling module (green) extracts latent motion features from videos with TPS transformations and synthesizes image frames; 2) the latent motion diffusion model (pink) generates motion features conditioned on the speech; 3) the refinement network (blue) restore missing details and produce the final fine-grained video.

rules is labor intensive. ANGIE [23] explicitly defines the problem of audio-driven co-speech gesture video generation, which utilizes an unsupervised feature, MRAA [35], to model body motion. Then a VQ-VAE [42] is leveraged to quantize common patterns, followed by a GPT-like network predicting discrete motion patterns to output gesture videos. However, as a coarse modeling of motion, MRAA is linear and fails to represent complex-shaped regions, limiting the quality of gesture videos generated by ANGIE. Differently, we carefully design a powerful latent motion feature and a matching generation model, enabling us to generate more realistic and stable gesture videos.

Conditional video generation. Another related task is conditional video generation. A variety of methods have been developed to generate videos conditioned on text [25], pose [18, 44], and also audio [30]. Recently, Diffusion models are used to model video space and exhibit promising results, but their computational requirements are often substantial due to the large volume of video data. Some works [14, 15, 33, 47] adopt an auto-encoder to create a latent space for videos and subsequently, diffusion generative models can focus solely on the latent space. However, these methods concentrate on generating general videos. The meaning of latent features is not well-defined, which may not always preserve the desirable information such as human motion. While LaMD [17] attempts to use two auto-encoders to separate content and motion, the separation is implicit and relies entirely on the design of the encoder network architecture. Additionally, the motion is represented as a vector without the time dimension, which may cause failure to model spatio-temporal variations in human gestures. In contrast, we design a time-aware diffusion model performing generation in a well-designed latent motion space tailored for gesture video generation and hence can generate gesture videos of high quality.

3. Our Approach

Given a speech audio $\boldsymbol{a}$ and a source image $\mathbf{S}$ of the speaker, our framework aims to generate an appropriate gesture video $\mathbf{V}$ (*i.e.* an image sequence). Due to the rich connotation of gesture videos, our overall concept is to decouple and generate motion information as a bridge in the video generation process. Therefore, the pipeline can be formulated as $\mathbf{V} = \mathcal{F}\left(\mathcal{G}\left(\mathcal{E}(\mathbf{S}), \boldsymbol{a}\right), \mathbf{S}\right)$, where $\mathcal{E}(\cdot)$ means motion decoupling to extract the source motion feature, which will be used with the audio as conditions to facilitate the audio-to-motion conversion by the diffusion model $\mathcal{G}(\cdot)$, and finally the image synthesis and refinement network $\mathcal{F}(\cdot)$ accomplish the refined motion-to-video generation.

In the following parts, we first explain the motion de-

coupling module with TPS transformation, which learns latent motion features from videos and guides the source image to warp to synthesize image frames containing desired gestures (Sec. 3.1). Then we elaborate the transformer-based diffusion model to perform generation within the latent motion space (Sec. 3.2). After that, we introduce the refinement network for better visual perception which focuses more on details of specific areas (Sec. 3.3). Finally, we present the inference process of the entire framework, where the optimal motion selection module helps to produce coherent and consistent long gesture videos (Sec. 3.4).

3.1. Motion Decoupling Module with TPS

To decouple human motion from videos, a straightforward method is to extract 2D poses with off-the-shelf pose estimators [8, 49]. However, as a zeroth-order model, poses completely discards appearance information around keypoints, making precise motion control and video rendering highly challenging. Furthermore, pre-training of pose estimators relies on hand-crafted annotations, suffering from error accumulation and often yielding jitters. The early work ANGIE [23] proposes to use MRAA [35] consisting of mean and covariance, which is linear and fails to model regions with intricate shapes. Besides, it is inappropriate to associate covariance directly with speech. Summarizing the above, we argue that an effective representation to decouple motion is crucial for the quality of generated gesture videos and their matching with speech. Therefore, we design a motion decoupling module based on a nonlinear transformation named TPS transformation, which deals well with curving edges and hence can model the motion of various-shaped body regions. Next, we will start by introducing TPS transformation as preliminary, followed by an exposition of the entire motion decoupling module.

TPS transformation. TPS transformation [5] aims to establish the mapping $\mathcal{T}_{tps}(\cdot)$ from the origin space $\mathbf{D}$ to the deformation space $\mathbf{S}$ by utilizing known paired keypoints as control, which takes the following form:

$$\begin{aligned} \mathcal{T}_{tps}(p) &= A \begin{bmatrix} p \\ 1 \end{bmatrix} + \sum_{i=1}^{N} w_i U\left(\left\|p_i^{\mathbf{D}} - p\right\|_2\right), \\ \text{s.t.} \quad & \mathcal{T}_{tps}(p_i^{\mathbf{D}}) = p_i^{\mathbf{S}}, \quad i = 1, 2, \ldots, N, \end{aligned} \tag{1}$$

where $p = (x, y)^\top$ denotes coordinate. $p_i^{\mathbf{D}}$ and $p_i^{\mathbf{S}}$ are the i^{th} paired keypoints from the origin and deformation space. $U(r) = r^2 \log r^2$ is a radial basis function. $A \in \mathbb{R}^{2\times 3}$ and $w_i \in \mathbb{R}^{2\times 1}$ are solvable parameters as introduced in [5].

In our setting, given a driving and a source image corresponding to the origin space $\mathbf{D}$ and the deformation space $\mathbf{S}$ separately, TPS transformation can establish local connections between the two frames, which will be further used to estimate a global optical flow $\mathcal{T}(\mathbf{D}) = \mathbf{S}$ [55]. It serves as the foundation for our motion decoupling module and offers two advantages: 1) as a flexible, non-linear transformation, it is suitable for modeling the motion of complex-shaped human bodies. 2) it relies solely on paired keypoints, whose movements are closely related to speech and thus can be more accurately controlled. Note that, unlike the keypoints of 2D poses only labeling certain joints, keypoints for TPS transformation come from adaptive boundary detection, involving both motion and crucial appearance information (*i.e.* region shapes), and can be easily used for operation at pixel level and further generating video frames.

The motion decoupling module is depicted as green in Fig. 2, which takes $\mathbf{S}$ and $\mathbf{D}$ as input, and outputs the constructed $\hat{\mathbf{D}}$ for end-to-end self-supervised training.

Keypoints predictor. To generate TPS transformation, we first design a keypoint predictor to predict $K \times N$ keypoints, which will subsequently be used for producing K TPS transformations with N points for each. The keypoints in $\mathbf{S}$ and $\mathbf{D}$ are estimated separately and then pairwise. The collection of keypoints $\{p_{ki}\}$ is very small in scale while being capable of generating a compact optical flow to animate images. So we take it as the latent motion feature.

Optical flow predictor. Now that we have K TPS transformations from predicted keypoint pairs modeling local motion, we can warp the source image $\mathbf{S}$ to obtain K deformed images. The optical flow predictor processes the stacked deformed images and finally outputs a pixel-level optical flow indicating global motion. Following [55], occlusion masks are also predicted, which will be fed into the image synthesis network together with the optical flow.

Image synthesis network. Due to misaligned pixels and occlusions in $\mathbf{S}$ and $\mathbf{D}$, direct warping fails to generate a valid reconstructed image $\hat{\mathbf{D}}$. Hence, we propose an image synthesis network of encoder-decoder architecture, with which $\mathbf{S}$ is encoded into feature maps in different scales. The warping operation is performed on these feature maps, and occlusion masks then guide them to be masked. Subsequently, the decoder synthesizes the constructed image $\hat{\mathbf{D}}$.

Training losses. From previous work [34, 35, 55], we use the perceptual construction loss, equivariance loss, and warping loss to train the whole module in an unsupervised manner. The final loss is the sum of the above:

$$\mathcal{L}_{tps} = \mathcal{L}_{per} + \mathcal{L}_{eq} + \mathcal{L}_{warp}. \tag{2}$$

For more details about training and the architecture, please refer to our supplementary material.

3.2. Latent Motion Diffusion Model

Since we have decoupled motion from gesture videos, our idea is to employ a diffusion model [16, 37, 38] for generation in the latent space by denoising pure Gaussian noise. Given a real video clip, we utilize the trained keypoint predictor to obtain the keypoint sets for all frames as

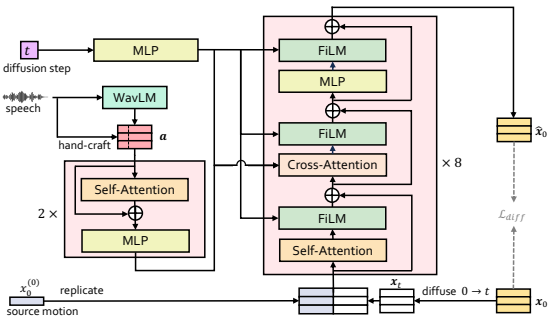

Figure 3. Noised motion features $\boldsymbol{x}_t$ are concatenated with replicated source $x_0^{(0)}$ and fed into our **transformer-based latent motion diffusion model** to predict the clean motion feature $\hat{\boldsymbol{x}}_0$ conditioned on the audio feature $\boldsymbol{a}$. The attention mechanism captures inherent connections between latent motion features and speech.

$\{p_{ki} \in \mathbb{R}^2\}^{(1:M)}$, where M is the frame number. We flatten the keypoints of each frame into a $C = K \times N \times 2$-dimensioned latent motion feature and finally get a feature sequence $\boldsymbol{x}_0 = x_0^{(1:M)} \in \mathbb{R}^{M \times C}$. Following [16], $\boldsymbol{x}_0$ will be diffused t times to get noised $\boldsymbol{x}_t$ and finally be cleaned.

Model. Per [28], our diffusion model predicts the clean motion feature sequence $\hat{\boldsymbol{x}}_0$ from noised $\boldsymbol{x}_t$ given noising step t and conditions $\boldsymbol{c} = \{\boldsymbol{a}, x_0^{(0)}\}$, where $\boldsymbol{a}$ denotes the audio feature, and $x_0^{(0)} \in \mathbb{R}^C$ is the source motion feature extracted from the source image $\mathbf{S}$, *i.e.* the first video frame.

During training, t is sampled from a uniform distribution $\mathcal{U}\{1, 2, \ldots, T\}$, and noised sequence $\boldsymbol{x}_t \in \mathbb{R}^{M \times C}$ is obtained by adding noise to $\boldsymbol{x}_0$ following DDPM [16]. Concerning speech audio features, [48] reveals that WavLM [9] features contain semantic information and are beneficial to the generation of co-speech motion. So we stack features generated from WavLM Large [9] with hand-crafted audio features to form a complete speech audio feature $\boldsymbol{a} \in \mathbb{R}^{M \times C_a}$. The former is interpolated to be aligned with the latter temporally, and $\boldsymbol{a}$ is also aligned with $\boldsymbol{x}_t$.

The latent motion diffusion model is in a transformer-like [40, 43] architecture as illustrated in Fig. 3, which is temporally aware and well-proven for modeling motion sequences [48]. The encoder takes the audio feature $\boldsymbol{a}$ as input and yields hidden speech embeddings. The decoder is a transformer decoder equipped with feature-wise linear modulation (FiLM) [26]. The source motion feature $x_0^{(0)}$ is replicated M times to have the same temporal length as $\boldsymbol{x}_t$, which are then concatenated together and fed into the self-attention network, capturing the temporal interactions within the motion sequence. After that, speech embeddings are projected to the cross-attention layer together with the output of self-attention, which facilitates learning the inherent relationship between the motion and speech sequence.

Training losses. We design the first term of loss to be common "simple" objective [16]. Besides, in the domain of motion generation, geometric losses [32, 39] are commonly used, which serve to constrain physical attributes and promote naturalness and coherence. Concerning the discussion in Sec. 3.1 that our latent features represent the motion, it is natural and reasonable to introduce geometric losses within the latent space. Here we use losses for velocity [36, 39] and acceleration [36]. The final training loss is as follows:

$$\mathcal{L}_{diff} = \mathcal{L}_{simple} + \lambda_{vel}\mathcal{L}_{vel} + \lambda_{acc}\mathcal{L}_{acc}. \tag{3}$$

Details can be found in the supplementary material.

3.3. Refinement Network

Guided by the motion features, the image synthesis network can generate speech-matched image frames according to the optical flow. However, we observe that the synthesized frames exhibit some blurs with missing details, especially in two types of regions: 1) occluded areas labeled by the occlusion masks, and 2) regions with complex textures such as hands and the face. As the image synthesis network is jointly trained with the motion decoupling module, to address this issue without disrupting the balance of motion modeling, we propose an independent refinement network.

We use a Unet-like architecture [29] equipped with residual blocks [52] to capture both global and local information. To draw more attention to occluded areas, the synthesized image frame is concatenated with the mask of the corresponding resolution mentioned in Sec. 3.1 and then fed into our refinement network. Additionally, in order to focus more on certain regions, we utilize MobileSAM [53] to segment hands and the face, and assign larger weights to both hands, face, and occluded areas in L1 reconstruction loss. Please refer to our supplementary material for more details.

3.4. Inference

As shown in Fig. 2, given a source image $\mathbf{S}$ and speech as inputs, keypoints of $\mathbf{S}$ are first detected with the keypoint predictor and gathered to form the source motion feature $x_0^{(0)}$. Conditioned on $x_0^{(0)}$ and extracted audio features $\boldsymbol{a}$, we randomly sample a Gaussian noise $\boldsymbol{x}_T \in \mathbb{R}^{M \times C}$ from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and denoise it via the DDPM reverse process. At each time step t, the denoised sample is predicted as $\hat{\boldsymbol{x}}_0 = \mathcal{G}(\boldsymbol{x}_t, t, \{\boldsymbol{a}, x_0^{(0)}\})$ and noised back to $\boldsymbol{x}_{t-1}$. After T steps, we obtain a clean sample $\boldsymbol{x}_0$. Repeating this procedure, we can get a consistent and coherent long sequence of motion features $\tilde{\boldsymbol{x}}$ with a novel optimal motion selection module, detailed further below. For each frame of $\tilde{\boldsymbol{x}}$, we can rearrange it to get $K \times N$ pairs of keypoints, producing K TPS transformations along with $x_0^{(0)}$ to estimate optical flow and occlusion masks. They are then fed into the image synthesis network to generate image frames, which will go through the refinement network together with corresponding masks and finally convert into fine-grained results. All frames gather to form a complete co-speech gesture video.

Optimal motion selection module. For the fact that meaningful co-speech gesture units last between 4-15 seconds [6, 46], it is crucial to generate motion feature sequences of any desired length. However, the transformer-based diffusion model, designed for fixed-length inputs, struggles with direct sampling of longer noise for generation due to both poor performance and high computational costs. A naive solution is to generate fixed-length segments for concatenation, where the source motion feature $x_0^{(0)}$ is replaced by the last frame of the previous segment to ensure continuity. However, in practice we notice that a single-frame condition cannot ensure the coherence and consistency between two segments, leading to flickers from position changes or jitters from direction changes of velocity.

To solve this problem, we propose an optimal motion selection module leveraging the diverse generative capability of the diffusion model, which operates solely at the inference stage. To be specific, from the second segment on, we generate P candidate sequences for the same audio segment. Then a lower-better score is calculated for each candidate according to two basic assumptions: within a small time interval of a real motion sequence, 1) keypoint positions are close; 2) keypoint velocity directions are similar. Finally, the candidate motion segment with the lowest score will be selected to extend the motion sequence. Details can be found in the supplementary material.

4. Experiments

4.1. Experimental Settings

Dataset and preprocessing. Data of our experiments is sourced from PATS dataset [1, 2, 13], consisting of transcribed poses with aligned audios and text transcriptions, containing around 84,000 clips from 25 speakers with a mean length of 10.7s, 251 hours in total. Similar to ANGIE [23], we perform our experiments on subsets of 4 speakers, including Jon, Kubinec, Oliver, and Seth. We download raw videos and audios to get clips according to PATS and conduct the following preprocessing steps: 1) Invalid clips with excessive audience applause, significant camera motion, or side views are excluded. 2) Clip lengths are limited to 4-15 seconds for meaningful gestures and resampled at 25 fps. 3) Frames are cropped with square bounding boxes, centering speakers, and resized to 256×256. 4) We extend these subsets with hand-crafted onset and chromagram features and WavLM [9] features. Finally, we obtain 1,200 valid clips for each speaker, randomly divided into 90% for training and 10% for evaluation, 4,800 in total.

Evaluation metrics. For motion-related metrics, we first extract 2D human poses with off-the-shelf pose estimator MMPose [31]. On this basis, we consider the quality, diversity, and alignment between gestures and speech, and choose: 1) **Fréchet Gesture Distance** (**FGD**) [51] to measure the distribution gap between real and generated gestures in the feature space, 2) **Diversity** (**Div.**) [24] which calculates feature distance between generated gestures on average. For these two metrics, we train an auto-encoder on poses from PATS. Also, we compute the average distance between closest speech beats and gesture beats as 3) **Beat Alignment Score** (**BAS**) following [21]. For video-related metrics, we utilize 4) **Fréchet Video Distance** (**FVD**) [41] to assess the overall quality of gesture videos. I3D [45] classifier pre-trained on Kinetics-400 [19] is used to compute FVD in the feature space.

4.2. Comparison to Existing Methods

We compare our method to: 1) the SOTA work ANGIE [23] in gesture video generation, and 2) MM-Diffusion [30], the SOTA work in video generation proven to be able to generate audio-driven human motion videos with experiments on AIST++ [21] human dance dataset.

The quantitative results are reported in Tab. 1. According to the comparison, our proposed approach significantly outperforms existing methods on motion-related metrics of FGD (56.44%) and Diversity (8.54%), which reveals that our motion-decoupled and diffusion-based generation framework is capable of generating realistic and diverse gestures in the motion space. Also, we achieve better performance on FVD than the best compared baseline MM-Diffusion, indicating that our method holds an advantage of ensuring the overall quality over the general audio-to-video method in gesture-specific settings. We notice that ANGIE with motion refinement tends to generate tremors synchronized with audio beat, leading to better results on BAS but at the expense of motion and visual quality. Fig. 4 presents frames of our generated videos compared with other methods, emphasizing the capacity of our method to generate videos with rich and realistic gestures matching the speech. On the contrary, limbs in ANGIE are modeled coarsely and vulnerable to abnormal deformations and absence from autoregressive error accumulation. MM-Diffusion struggles to capture body structures, leading to more or no hands.

Additionally, owing to the capability of TPS transformation to model complex-shaped regions and the close association between motion and speech established by the diffusion model, our method excels in generating precise and diverse fine-grained hand movements. As shown in Fig. 5, directly generated videos by MM-Diffusion entirely fail to produce reasonable hand morphology. While ANGIE attempts to utilize MRAA to represent motion, this linear affine transformation coarsely models curved body regions with Gaussian distribution, resulting in hand movements presented as the translation (controlled by the mean), rotation and scaling (controlled by PCA parameters of the covariance) of an "ellipse" in ANGIE's results. In contrast, our method generates hand movements matching the

Table 1. Quantitative results on test set. Bold indicates the best and underline indicates the second. For ANGIE [23] we reproduce the code. For MM-Diffusion [30] we use the officially published code. Subjective evaluation is results of MOS with 95% confidence intervals.

Name	Objective evaluation				Subjective evaluation			
	FGD ↓	Div. ↑	BAS ↑	FVD ↓	Realness ↑	Diversity ↑	Synchrony ↑	Overall quality ↑
Ground Truth (GT)	8.976	5.911	0.1506	1852.86	4.76±0.05	4.70±0.06	4.77±0.05	4.73±0.06
ANGIE	55.655	5.089	**0.1504**	2965.29	2.07±0.08	2.53±0.08	2.19±0.08	2.00±0.07
MM-Diffusion	41.626	5.189	0.1098	2656.06	1.77±0.08	2.02±0.09	1.69±0.08	1.47±0.07
Ours	**18.131**	**5.632**	0.1273	**2058.19**	**3.79±0.08**	**3.91±0.07**	**3.90±0.08**	**3.77±0.07**

Table 2. Ablation study results. Bold indicates the best and underline indicates the second. 'w/o' is short for 'without'.

Name	Objective evaluation				Subjective evaluation			
	FGD ↓	Div. ↑	BAS ↑	FVD ↓	Realness ↑	Diversity ↑	Synchrony ↑	Overall quality ↑
w/o TPS + MRAA	288.378	4.625	0.1200	3034.71	2.59±0.09	2.50±0.09	2.59±0.09	1.96±0.07
w/o WavLM	37.072	5.344	0.1253	**2053.44**	3.44±0.08	3.45±0.08	3.43±0.08	3.38±0.07
w/o refinement	26.125	5.549	**0.1288**	2154.00	3.67±0.08	3.75±0.08	3.74±0.07	3.49±0.06
LN Samp.	46.055	4.871	0.1250	2236.72	2.65±0.09	2.25±0.09	2.45±0.09	2.70±0.09
Concat.	20.964	5.596	0.1250	2085.50	3.66±0.07	3.64±0.08	3.71±0.08	3.67±0.07
Ours	**18.131**	**5.632**	0.1273	2058.19	**3.79±0.08**	**3.91±0.07**	**3.90±0.08**	**3.77±0.07**

speech, featuring intricate and plausible variations in hand shapes, which is crucial for high-quality human gestures.

User study. In practice, objective metrics may not always be consistent with human subjective perceptions [48], especially in the novel setting of co-speech gesture video generation. To gain further insights into the visual performance of our method, we conduct a user study to evaluate the gesture videos generated by each method alongside the ground truth. For each method, we sample 24 generated videos from the PATS test set between 3.2-12.8 seconds. 20 participants are invited to conduct the Mean Opinion Scores (MOS). Participants are asked to rate the videos in four aspects: 1) **Realness**, 2) **Diversity**, 3) **Synchrony** between speech and gestures, and 4) **Overall quality**. The first three focus on motion, while the last places more emphasis on visual perceptions. The rating scale ranges from 1 to 5 with a 1-point interval, where 1 means the poorest and 5 means the best. The results are reported in the last four columns in Tab. 1. Our method significantly surpasses other methods in all dimensions, which reveals that our framework can generate better gesture videos considering both motion and overall visual effects. It is noteworthy that the slight advantage of ANGIE on BAS does not translate into better gesture-speech synchrony in human subjective evaluation, where excessive tremors are not considered in sync with the speech. Please refer to the supplement for the effectiveness and robustness analysis of BAS and other objective metrics. According to the feedback from participants, "before seeing the ground truth", our generated gesture videos are already "natural and well-matched to the speech enough to be mistaken as real". Besides, there is an interesting finding that despite our emphasis on excluding irrelevant factors like textures and facial expressions in motion-related evaluations, participants express that "when compared with the ground truth containing rich details, although generated motion is realistic, they are inevitably influenced by appearance factors". This demonstrates that human perception of motion and appearance are interrelated. Hence generating co-speech gesture videos with visual appearance is a meaningful problem in the field of human-machine interaction.

4.3. Ablation Study

We conduct an ablation study to demonstrate the effectiveness of different components in our framework. The results are shown in Tab. 2. We explore the effectiveness of the following components: 1) the TPS-based motion decoupling module, 2) WavLM features, 3) the refinement network, and 4) the optimal motion selection module.

Supported by the results in Tab. 2, when we replace TPS-based motion features with MRAA following ANGIE, FGD and FVD severely deteriorate by 1490% and 47.4%. When WavLM features are removed, FGD, Diversity, and BAS all deteriorate for the fact that WavLM features contain rich high-level information like semantics and emotions, crucial for driving gestures. However, WavLM brings a slight increase in FVD by 4.75, although not significantly (0.23%), demonstrating that the positive impact of WavLM is evident in motion while having subtle effects in visual aspects. The refinement network brings improvements in FGD, Diversity, and FVD, especially FVD decreased by 95.81 (4.4%). Detailed analysis and visual comparisons of our ablation study can be found in the supplementary material.

For the optimal motion selection module, we replace it with two simple strategies to generate longer videos as men-

Figure 4. **Visual comparison with SOTAs.** Our method generates gestures with a broader range of motion (dashed boxes) matching both beats (green words) and semantics (purple words). Red boxes denote unrealistic gestures generated by ANGIE [23] and MM-Diffusion [30].

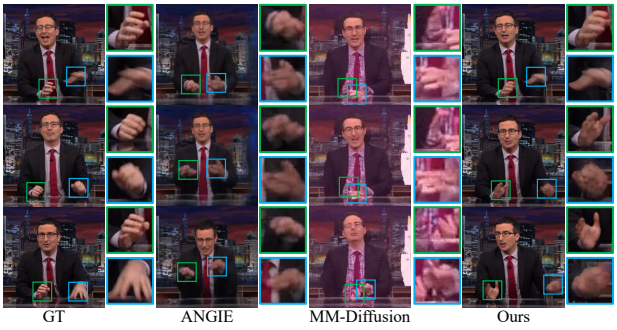

Figure 5. **Visualization results of fine-grained hand variations.** Our generated gesture videos are more plausible and diverse.

tioned in Sec. 3.4: 1) long noise sampling (LN Samp.), and 2) direct concatenation (Concat.). According to Tab. 2, our method equipped with the optimal motion selection module achieves the best performance across all dimensions.

User study. Similarly, we conduct a user study for ablations as described in Sec. 4.2. Results in Tab. 2 indicate that the final performance of our model decreases without any module. Consistent with our expectations, removing TPS has the most significant impact on the results of Realness. This reiterates the crucial significance of employing an appropriate motion feature to decouple motion. Besides, we also conduct another user study in the context of longer video generation and report the results in the supplement.

5. Conclusion

In this paper, we present a novel motion-decoupled framework for co-speech gesture video generation without structural human priors. Specifically, we carefully design a nonlinear TPS transformation to obtain latent motion features, which describe motion trajectories while retaining crucial appearance information. Then, a transformer-based diffusion model is used within this latent motion space to model the intricate temporal relationship between gestures and speech, followed by an optimal motion selection module to generate diverse long gesture videos. Besides, a refinement network is leveraged to draw more attention to certain details and bring better visual effects. Extensive experiments demonstrate that our framework produces long-term realistic, diverse gesture videos appropriate to the given speech, and significantly outperforms existing approaches.

Acknowledgments

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

References

[1] Chaitanya Ahuja, Dong Won Lee, Ryo Ishii, and Louis-Philippe Morency. No gestures left behind: Learning relationships between spoken language and freeform gestures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1884–1895, 2020. 2, 6

[2] 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 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 248–265. Springer, 2020. 2, 6

[3] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. Style-controllable speech-driven gesture synthesis using normalising flows. In *Computer Graphics Forum*, pages 487–496. Wiley Online Library, 2020. 2

[4] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8340–8348, 2018. 1

[5] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6):567–585, 1989. 2, 4

[6] Peter Bull. Gesture: Visible action as utterance. *Journal of Language and Social Psychology*, 25(3):339–341, 2006. 6

[7] Judee K Burgoon, Thomas Birk, and Michael Pfau. Nonverbal behaviors, persuasion, and credibility. *Human communication research*, 17(1):140–169, 1990. 1

[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 4

[9] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16 (6):1505–1518, 2022. 5, 6

[10] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018. 2

[11] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*, pages 102–118. Springer, 2022. 2

[12] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F Troje, and Marc-André Carbonneau. Zeroeggs: Zero-shot example-based gesture generation from speech. In *Computer Graphics Forum*, pages 206–216. Wiley Online Library, 2023. 2

[13] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019. 1, 6

[14] Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin Minaee, Dimitris Metaxas, and Sergey Tulyakov. Show me what and tell me how: Video synthesis via multimodal conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3615–3625, 2022. 2, 3

[15] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 2, 3

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 4, 5

[17] Yaosi Hu, Zhenzhong Chen, and Chong Luo. Lamd: Latent motion diffusion for video generation. *arXiv preprint arXiv:2304.11603*, 2023. 3

[18] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. *arXiv preprint arXiv:2304.06025*, 2023. 2, 3

[19] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[21] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation. *arXiv preprint arXiv:2101.08779*, 2(3), 2021. 6

[22] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European Conference on Computer Vision*, pages 612–630. Springer, 2022. 2

[23] Xian Liu, Qianyi Wu, Hang Zhou, Yuanqi Du, Wayne Wu, Dahua Lin, and Ziwei Liu. Audio-driven co-speech gesture video generation. *Advances in Neural Information Processing Systems*, 35:21386–21399, 2022. 1, 3, 4, 6, 7, 8

[24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10462–10472, 2022. 2, 6

[25] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18444–18455, 2023. 3

[26] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a

general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 5

[27] Shenhan Qian, Zhi Tu, Yihao Zhi, Wen Liu, and Shenghua Gao. Speech drives templates: Co-speech gesture synthesis with learned templates. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11077–11086, 2021. 1

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 5

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2, 5

[30] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10219–10228, 2023. 2, 3, 6, 7, 8

[31] Arindam Sengupta, Feng Jin, Renyuan Zhang, and Siyang Cao. mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns. *IEEE Sensors Journal*, 20 (17):10032–10044, 2020. 6

[32] Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *ACM Transactions on Graphics (TOG)*, 40(1):1–15, 2020. 5

[33] Gaurav Shrivastava and Abhinav Shrivastava. Diverse video generation using a gaussian process trigger. *arXiv preprint arXiv:2107.04619*, 2021. 2, 3

[34] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019. 4

[35] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662, 2021. 3, 4

[36] Li Siyao, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11050–11059, 2022. 5

[37] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 4

[38] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020. 4

[39] 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. 2, 5

[40] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023. 5

[41] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6

[42] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[43] 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. 5

[44] Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for referring human dance generation in real world. *arXiv preprint arXiv:2307.00040*, 2023. 2, 3

[45] Xianyuan Wang, Zhenjiang Miao, Ruyi Zhang, and Shanshan Hao. I3d-lstm: A new model for human action recognition. In *IOP Conference Series: Materials Science and Engineering*, page 032035. IOP Publishing, 2019. 6

[46] Jason R Wilson, Nah Young Lee, Annie Saechao, Sharon Hershenson, Matthias Scheutz, and Linda Tickle-Degnen. Hand gestures and verbal acknowledgments improve human-robot rapport. In *Social Robotics: 9th International Conference, ICSR 2017, Tsukuba, Japan, November 22-24, 2017, Proceedings 9*, pages 334–344. Springer, 2017. 1, 6

[47] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 2, 3

[48] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. *arXiv preprint arXiv:2305.04919*, 2023. 2, 5, 7

[49] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 4

[50] 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 *2019 International Conference on Robotics and Automation (ICRA)*, pages 4303–4309. IEEE, 2019. 2

[51] 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. 6

[52] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 2, 5

[53] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023. 5

[54] 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. 2

[55] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022. 2, 4

[56] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 2

[57] Yang Zhou, Jimei Yang, Dingzeyu Li, Jun Saito, Deepali Aneja, and Evangelos Kalogerakis. Audio-driven neural gesture reenactment with video motion graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3418–3428, 2022. 2