DR^2: Disentangled Recurrent Representation Learning for Data-efficient Speech Video Synthesis

Chenxu Zhang[1], Chao Wang[1], Yifan Zhao[2], Shuo Cheng[3], Linjie Luo[1], Xiaohu Guo[4]
[1]ByteDance Inc [2]Peking University
[3]Georgia Institute of Technology [4]The University of Texas at Dallas
{chenxuzhang, chao.wang, linjie.luo}@bytedance.com, zhaoyf@pku.edu.cn,
shuocheng@gatech.edu, xguo@utdallas.edu

Abstract

Although substantial progress has been made in audio-driven talking video synthesis, there still remain two major difficulties: existing works 1) need a long sequence of training dataset (>1h) to synthesize co-speech gestures, which causes a significant limitation on their applicability; 2) usually fail to generate long sequences, or can only generate long sequences without enough diversity. To solve these challenges, we propose a Disentangled Recurrent Representation Learning framework to synthesize long diversified gesture sequences with a short training video of around 2 minutes. In our framework, we first make a disentangled latent space assumption to encourage unpaired audio and pose combinations, which results in diverse "one-to-many" mappings in pose generation. Next, we apply a recurrent inference module to feed back the last generation as initial guidance to the next phase, enhancing the long-term video generation of full continuity and diversity. Comprehensive experimental results verify that our model can generate realistic synchronized full-body talking videos with training data efficiency.

1. Introduction

Generating realistic human speech video from input audio is a long-standing objective in computer graphics and computer vision with key applications to virtual humans. Many existing works focused on speech video generation with a head part only [14, 64, 66] or head and shoulder together [24, 54, 68], while others extended to the synthesis of body gestures as well, thus increasing the overall expressiveness of the results. Among these works, early methods utilize heuristics and rules to generate gestures triggered from a predefined audio-to-gesture mapping [10, 44, 45]. This will, however, result in repetitive gestures and a lack of person-specific idiosyncrasies given similar input audio.

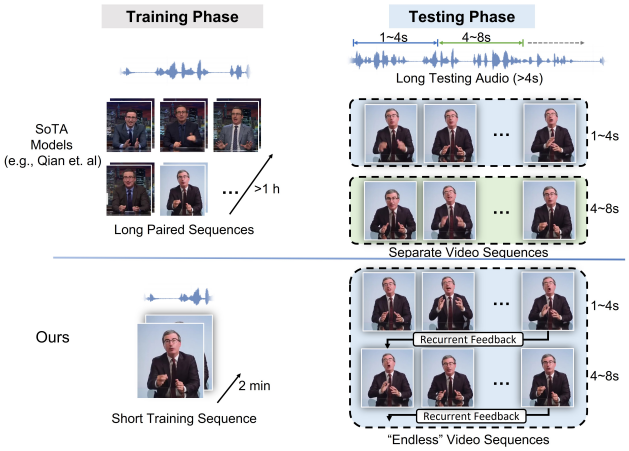

Figure 1. Comparison illustration of state-of-the-art and our method. 1) Training phase: existing methods require long training data (>60 mins) while our method can be applied with **only 2 mins short videos**. 2) Testing phase: existing methods only consider the diversity of short video fragments (every 4s), while our method can generate **endless video sequences** with high diversity and continuity.

Recent works [23, 38, 42, 49] apply data-driven approaches to predict more diverse gestures by learning human speech behaviors from collected data. However, their applicability is significantly limited by: 1) large training data needed, and 2) less diversity in long sequence generation:

- In the training phase, prevailing methods require long-term training sequences (>1h) of the same actor [23, 49] or multi-actors with similar gestures [36, 42] as shown in Fig. 1. It creates a challenge for common applications where it is usually infeasible to collect sufficient data for one specific actor. In addition to the difficulty in data collection, speech gestures vary in terms of speech content, camera position, and speech scenes, spanning over sitting, standing, and moving scenarios. Fitting models with monotonous training data could

Table 1. Comparisons of state-of-the-art speech synthesis methods and our proposed method.

Methods	Short Training Dataset	Facial Expression	Realistic Video	Short Diversified Gesture	Long Diversified Gesture
Speech2Gesture [23]	✗	✗	✓	✗	✗
SDT [49]	✗	✓	✓	✓	✗
Audio2Gesture [36]	✗	✗	✗	✓	✗
HA2G [42]	✗	✗	✗	✓	✗
Ours	✓	✓	✓	✓	✓

lead to less diversity as shown in Tab. 1.

- In the testing phase, earlier adversarial learning models [23] take input audio as guidance and generate the corresponding gestures. In addition to the difficulty of training, these models fail to capture the "one-to-many" mapping of audio-to-gestures. As shown in Fig. 1, recent ideas to solve this [36, 49] take short sequenced gestures (*e.g.*, 4s) beyond audio as input for generation. However, these methods fail to generate long diversified sequences. The reason is that their gesture features only represent body motion in a few seconds, and only the same gesture features could appear in each test, which leads to fragmented generation with possible repetition among short sequences and discontinuity between two adjacent sequences.

In this paper, we aim to answer this question: how to generate *long-term* speech video with continuity and diversity by using only *short-term* training data? We propose a novel Disentangled Recurrent Representation (DR^2) learning mechanism for this data-efficient speech video synthesis task. Prevailing models aim to embed input audio and pose sequences into a unified latent space while learning a "one-to-one" mapping of audio and poses. Unlike this learning framework, we disentangle the learning of audio and poses and use only one-shot initial pose and input guidance, leaving the network a rich imagined space to disentangle audio and pose features. Besides the paired learning of audio and poses, we construct a state bank to store feasible initial poses but different from the paired one. During the training phase, we randomly select the unpaired pose and audio features to encourage the diversity of "one-to-many" mapping. Benefiting from the one-shot pose state, we develop a recurrent inference module for synthesizing arbitrarily-long sequences. In Fig. 1, the pose state generated by the last sequence serves as the *initial prior* and *gesture template* to guide the generation of the following sequence, resulting in long-term diverse videos with continuity.

In our disentangled recurrent learning framework, we resort to SMPL-X [46] model as the intermediate 3D representation for constructing physiologically reasonable models, which avoids unnatural deformations in conventional keypoint-based representations. Beyond the basic representation in previous methods, we extend the generation objective by incorporating 3D hand embedding and extend the application scenarios to sitting, standing, and moving status. Finally, we develop a neural rendering network to synthesize these 3D models into vivid 2D videos. Comprehensive evaluations on both public and our self-captured datasets demonstrate the effectiveness of our method.

The main contributions are as follows: 1) We propose a novel Disentangled Recurrent Representation (DR^2) learning framework to synthesize arbitrarily-long diversified gesture sequences from short training videos of around 2 minutes only. 2) We propose a disentangled module with a state bank to encourage the learning of *unpaired pose* and *audio embedding*, resulting in diverse "one-to-many" mappings in pose generation. 3) We develop a recurrent inference module to feed back the last generation as *initial pose prior* and *gesture template*, which determines both the start pose and the general appearance of the next sequence, resulting in arbitrarily-long diverse poses with full continuity.

2. Related Work

Audio/Text-driven gesture synthesis. Traditional rule-based gesture generation methods [10, 11, 44, 45] are always phoneme-dependent and restricted to special language-specific rules. However, early data-driven methods find appropriate audio-to-gesture mapping based on statistical modeling techniques, such as Markov models [9, 34, 35], conditional Restricted Boltzmann Machine [16], conditional random fields [17] and dynamic Bayesian Network [50]. Recent deep-learning techniques have started to be widely used in gesture synthesis. Text-driven gesture prediction usually builds a mapping between text semantic information to co-speech gestures [2, 8, 29, 63], while audio-driven gesture generation methods usually predict 2D or 3D gestures using diverse neural network architecture base such as RNN [32], GRU [60], GAN [1, 7, 25, 26], LSTM [3, 21, 28, 51, 53]. Audio2gesture [36] split the VAE latent code of motion gestures into shared code and motion-specific code to regress training data and generate diverse motions.

Other methods [70] exploit multi-modal information to improve the correlation between generated gestures and other modalities, including audio, text transcripts, speaker identities, styles, and expressions [4, 33, 37, 40–42, 61, 62]. Gesticulator [33] generates beat and semantic gestures together from audio and text input, and Ao *et al.* [4] extend this by regularizing the rhythmically and semantically consistency. Ghorbani *et al.* propose ZeroEGGS [22] to generate gestures with zero-shot style control and a new multi-modal dataset with multiple styles. Liu *et al.* present DisCo [39] that captures both high and low-frequency occurrence information with more diverse motions by disentangling motion data into implicit content and rhythm features. Yang *et al.* [59] synthesize natural co-speech gestures due to its novel quantization-based and phase-guided motion matching framework. The other line of work [5] utilizes diffusion models to generate stylized co-speech gestures with flexible style control. Besides, Zhu *et al.* [69] utilize a Diffusion Audio-Gesture Transformer to obtain coherent gestures with better mode coverage and stronger audio correlations. Yang *et al.* [58] introduce cross-local attention and self-attention models to the gesture diffusion pipeline to obtain stylized and diverse gestures. These models rely on large-scale training datasets, such as [39, 59] that use disentanglement to enrich audio information with additional features, aiming to prevent overly smoothed results when training on large datasets. However, they still face difficulties with short training datasets. In this manner, we use the disentanglement method with paired and unpaired training to address data-efficiency issues.

Audio-driven talking video generation. Many recent researchers start to apply audio-driven face expression and gesture synthesis techniques in realistic talking video generation [23, 38, 46, 49]. One type method is the face generation [14, 18, 20, 64, 66, 67]. Another category of methods work on the head and shoulders (upper torso) together [13, 24, 43, 47, 52, 54, 65, 68]. So far only a few audio-to-gesture works [23, 38, 49] are focusing on generating half-body talking videos. Speech2gesture [23] first leverages a UNet-based framework to predict 2D half-body gestures, and then utilizes a video-to-video network [12] to synthesize the final talking video. Liao *et al.* [38] trained an LSTM-based network to predict SMPL-X parameters [46], whose 3D joints information is further fed into a video-to-video network [55] to synthesize final output videos. SDT [49] learns a set of gesture template vectors to control the style of 2D gestures, and uses an image warping and translation module [6] to generate final synthesized talking images per frame. Note that all these previous works are not data-efficient, and generate long gesture sequences with repetitive patterns. In contrast, our method only needs a shorter training sequence and generates diverse long sequences of gestures and the corresponding talking videos (Tab. 1).

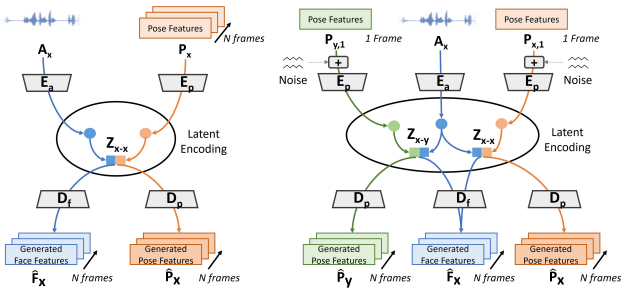

Figure 2. Comparison of conventional learning and our disentangled representation assumptions. (Left) Conventional methods learn the generation from paired audio and pose encoding. (Right) Our method assumes the pose and audio features of related clips can be disentangled and composed into new diverse generations.

3. The Method

3.1. Assumption

Given an input $\mathbf{V}_x \sim \mathcal{V}$ (the xth video clip in the training data), we denote the paired audio-gesture state as $(\mathbf{A}_x, \mathbf{P}_x) \in \mathbf{V}_x$. As shown in Fig. 2 (Left), conventional video synthesis methods assume that the audio features and pose gestures should be encoded in the same latent space to get a fused embedding $\mathbf{Z}_{x-x} = [E_a(\{\mathbf{A}_{x,i}\}_{i=1}^n), E_p(\{\mathbf{P}_{x,i}\}_{i=1}^n)]$, where E_a and E_p are audio and pose encoders, n represents the number of frames in a clip, and $\mathbf{A}_{x,i}$ and $\mathbf{P}_{x,i}$ denote the ith frame audio and pose in each clip respectively. From this latent space, the encoding $\mathbf{Z}_{x-x}$ are further decoded by D_p and D_f to generate the synthesized pose sequences $\hat{\mathbf{P}}_x$ and face sequences $\hat{\mathbf{F}}_x$. However, this prevailing learning framework [23, 49] requires a long sequence of training data with similar identities and would lead to overfitting due to the strong prior encoded by the gesture sequences.

As the gesture generation of speech video synthesis is an ill-posed problem, *i.e.*, one can perform different gestures with the same speech content, learning a strict one-to-one mapping in prevailing works encounters challenges whenever there is no sufficient data collected. Toward this end, we design a *disentangled latent space* for gesture and audio combinations to embed this loose correlation. As shown in Fig. 2 (Right), our method works in a one-shot fashion, which only adopts the first pose state $\mathbf{P}_{x,1}$ as initial input while leaving the network a huge imagined space to fill in the subsequent ones $\mathbf{P}_{x,2:n}$. Let $\mathbf{V}_y$ be another clip sampled from the same video space $\mathcal{V}$ and $\mathbf{P}_y$ be the gesture state in $\mathbf{V}_y$. Different from existing methods, we then use the one-shot guidance with random noise $\xi \sim \mathcal{N}(0, \sigma)$ to obtain: (1) the paired latent encoding $\mathbf{Z}_{x-x} = [E_a(\{\mathbf{A}_{x,i}\}_{i=1}^n, E_p(\mathbf{P}_{x,1} + \xi)]$; and (2) the unpaired encoding $\mathbf{Z}_{x-y} = [E_a(\{\mathbf{A}_{x,i}\}_{i=1}^n, E_p(\mathbf{P}_{y,1} + \xi)]$.

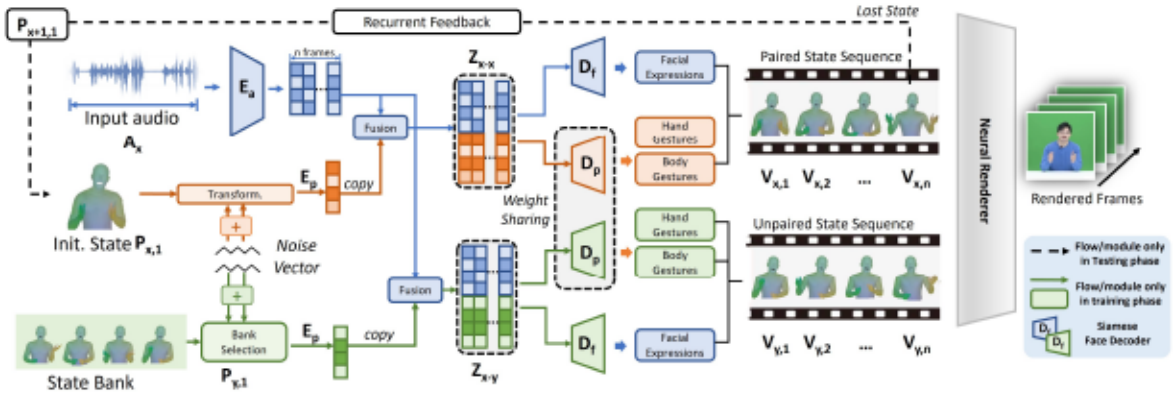

Figure 3. Pipeline of our disentangled recurrent representation learning framework. Given an input audio $\mathbf{A}_x$ and an initial pose state $\mathbf{P}_{x,1}$ as guidance, our network generates a paired face, hand and body embedding as the basic supervision. The second stream encodes the sample audio $\mathbf{A}_x$ but with one different initial state $\mathbf{P}_{y,1}$ to compose into an unpaired generation. During the inference phase, we feed back the last pose state $\mathbf{P}_{x,n} \in \mathbf{V}_{x,n}$ as the initialization $\mathbf{P}_{x+1,1}$ of the next generation.

We assume that $\mathbf{Z}_{x-y}$ is also a reasonable latent encoding and can be decoded into complete pose and face sequences, $\{\hat{\mathbf{F}}_{x,t}\}_{t=1}^{n} = D_f(\mathbf{Z}_{x-y}), \{\hat{\mathbf{P}}_{y,t}\}_{t=1}^{n} = D_p(\mathbf{Z}_{x-y})$. Note that we use the weight-sharing pose decoder D_p for paired $x-x$ and unpaired $x-y$ latent encodings. Thus our designed framework looses the coupling of the audio and gesture from the same video frame while encouraging the diversity of poses in the video space $\mathcal{V}$.

3.2. Disentangled Recurrent Representation

Framework. Based on this idea of disentangled latent space, we first build a paired construction module as in the top half of Fig. 3. With the audio encoder E_a of all n frames, we first use the initial pose state $\mathbf{P}_{x,1}$ with a random noise ξ to build a self-supervised output of pose gestures $\{\hat{\mathbf{P}}_{x,t}\}_{t=1}^{n}$ and facial expressions $\{\hat{\mathbf{F}}_{x,t}\}_{t=1}^{n}$ for all n frames. For the unpaired training branch in the lower half of Fig. 3, we first construct a state bank $\mathcal{B}$ to store all initial states from the unpaired poses, and then select an unpaired one $\forall \mathbf{P}_{y,1} \in \mathcal{B}, \mathbf{P}_{y,1} \neq \mathbf{P}_{x,1}$. Analogous to the paired training, we then replicate $\mathbf{P}_{y,1}$ features by n times to fuse with the audio feature $\mathbf{A}_x$. After that, the encoded feature $\mathbf{Z}_{x-x}$ (paired) and $\mathbf{Z}_{x-y}$ (unpaired) are decoded separately by two decoders: the pose decoder D_p to generate hand and body gestures, and face decoder D_f to generate facial expressions. Note that the general appearance (body and hand gestures) shows a strict correlation with the initial pose state while the gestures' movements have a relatively high correlation with the audio input to encourage diversity. With the limitation of short training data (less than 1/30 of other datasets), our disentangled representation learning is specifically designed for such constraints: 1) we do not input a full pose sequence of n frames but only the initial frame to encourage the strict correlation with initial pose; 2) we loose the coupling of audio and pose and use the unpaired training to enhance the diversity of poses; 3) the random noises are used to add a slight perturbation for initial states while still maintaining them in a reasonable range.

Recurrent inference for arbitrarily-long sequences. Existing works encounter great challenges when encoding and generating long-term diverse sequences. As shown in Fig. 2 (Left), these methods propose to encode a certain length of poses *e.g.*, a clip of 128 frames, and to decode and generate an output of 128 frames. This encoding-decoding framework would lead to two problems (Especially for short dataset): 1) Input pose sequences strictly constrain the generation results to be similar. However, the training and inference audios usually show a huge distribution gap, which makes them hard to generalize. 2) The generated sequence can be consistent within each clip, but it is hard to guarantee the continuity between two adjacent clips since they are generated separately. To address this, we propose recurrent feedback of the pose state as shown in Fig. 3. At the t-th clip our framework generates $\{\hat{\mathbf{P}}_{x,t}\}_{t=1}^{n}(t)$, and in the $(t+1)$-th clip, we have:

$$\begin{aligned} &\{\hat{\mathbf{P}}_{x,t}\}_{t=1}^{n}(t) = D_p(E_a(\{\mathbf{A}_{x,t}(t)\}_{t=1}^{n}), E_p(\mathbf{P}_{x,1}(t))), \\ &\mathbf{P}_{x,1}(t+1) \leftarrow \hat{\mathbf{P}}_{x,n}(t). \end{aligned} \tag{1}$$

With such recurrent scheme, the generated ending pose of the current clip is used as an initial guidance and gesture template of the next clip. The random noise ξ is removed for sequences continuity during inference stage. Thus we can generate arbitrarily-long diverse video sequences. Besides, as we only encode the pose of first frame as guidance, our generated video shows rich pose diversity.

3.3. Models and Learning Objective

3D-aware audio2gesture network. Directly predicting 3D joints or 2D keypoints without articulation constraints would lead to physiologically unreasonable movements in gesture generation. Hence we use a 3D parametric model SMPL-X [46] as the intermediate representation to regularize the movement of body parts. As shown in Fig. 3, given an input audio $\mathbf{A}_x$, we first use DeepSpeech [27] as feature extractor $E_a(\{\mathbf{A}_{x,t}\}_{t=1}^n) \in \mathbb{R}^{29\times n}$ to encode it into latent space. We use SMPL-X to represent 3D pose, which is composed of a body gesture of $\hat{\mathbf{P}}_{body} \in \mathbb{R}^{35\times n}$ and a hand gesture of $\hat{\mathbf{P}}_{hand} \in \mathbb{R}^{24\times n}$. For simplicity, they are noted as *pose gestures* in this paper. The facial decoder D_f generates a face expression of $\hat{\mathbf{F}} \in \mathbb{R}^{10\times n}$. The SMPL-X model is then applied to map these parameters to its skin mesh vertices $\mathbf{M} \in \mathbb{R}^{10475\times 3\times n}$ and joints $\mathbf{J} \in \mathbb{R}^{127\times 3\times n}$.

Learning objectives. After fitting the training video with SMPL-X, we can obtain the ground-truth face expressions $\mathbf{F}^G$, and hand-body poses $\mathbf{P}^G$. For frame t in video $\mathbf{V}_x$, we use $\hat{\mathbf{F}}_{x,t} \in \mathbb{R}^{10}$ and $\hat{\mathbf{P}}_{x,t} \in \mathbb{R}^{59}$ to represent the predicted facial expressions and pose gestures respectively. Random noise $\xi \sim \mathcal{N}(0, \sigma)$ is applied to both $\mathbf{P}_x^G$ and $\mathbf{P}_y^G$ to enhance the diversity of generated sequences. Please refer to our supplementary for more details.

Our learning objective comprises three parts: 1) initial state reconstruction; 2) paired sequence reconstruction; 3) unpaired sequence reconstruction. The initial pose reconstruction follows the standard MSE constraint, which aims to guarantee the stable initialization of each generated clip. $\forall\, \mathbf{V}_x \in \mathcal{V}$:

$$\mathcal{L}_x^{init} = \|\hat{\mathbf{P}}_{x,1} - \mathbf{P}_{x,1}^G\|_2^2 + \|\hat{\mathbf{F}}_{x,1} - \mathbf{F}_{x,1}^G\|_2^2. \tag{2}$$

The paired sequence loss of poses consists of two terms, which regress the pose sequence reconstruction and maintain the motion between adjacent frames to be similar:

$$\begin{aligned}\mathcal{L}_{x-x}^{pair}(\mathbf{P}) &= \|D_p(E_p(\mathbf{P}_{x,1}^G), E_a(\{\mathbf{A}_{x,t}\}_1^n)) - \mathbf{P}_x^G\|_2^2 \\ &+ \frac{\lambda_p}{n-1}\sum_{t=1}^{n-1}\|(\hat{\mathbf{P}}_{x,t+1} - \hat{\mathbf{P}}_{x,t}) - (\mathbf{P}_{x,t+1}^G - \mathbf{P}_{x,t}^G)\|_2^2.\end{aligned} \tag{3}$$

Similarly, the paired sequence loss of facial expressions is defined as:

$$\begin{aligned}\mathcal{L}_{x-x}^{pair}(\mathbf{F}) &= \|D_f(E_p(\mathbf{P}_{x,1}^G), E_a(\{\mathbf{A}_{x,t}\}_1^n)) - \mathbf{F}_x^G\|_2^2 \\ &+ \frac{\lambda_f}{n-1}\sum_{t=1}^{n-1}\|(\hat{\mathbf{F}}_{x,t+1} - \hat{\mathbf{F}}_{x,t}) - (\mathbf{F}_{x,t+1}^G - \mathbf{F}_{x,t}^G)\|_2^2.\end{aligned} \tag{4}$$

Besides the paired training, our disentangled representation learning constructs unpaired constraints by selecting another video clip from the state bank. $\forall\, (\mathbf{V}_x, \mathbf{V}_y) \in \mathcal{V}, \mathbf{V}_x \neq \mathbf{V}_y$. The overall motion and the detailed per-frame movements are further disentangled for unpaired gesture learning:

$$\begin{aligned}\mathcal{L}_{x-y}^{un}(\mathbf{P}) &= \|D_p(E_p(\mathbf{P}_{y,1}^G), E_a(\{\mathbf{A}_{x,t}\}_1^n)) - \mathbf{P}_y^G\|_2^2 \\ &+ \frac{\lambda_p}{n-1}\sum_{t=1}^{n-1}\|(\hat{\mathbf{P}}_{y,t+1} - \hat{\mathbf{P}}_{y,t}) - (\mathbf{P}_{x,t+1}^G - \mathbf{P}_{x,t}^G)\|_2^2.\end{aligned} \tag{5}$$

The first term of Eq. (5) is a reconstruction loss that contributes to the overall motion of the whole sequence conditioned on the initial unpaired pose prior $\mathbf{P}_{y,1}^G$. The second term is a frame motion loss that contributes to the detailed movements (*e.g.*, related to audio beats) and is learned by the velocity of the paired $\mathbf{P}_x^G$ to maintain the multi-modal consistency, indicating a similar motion velocity generated from the same audio beat.

The unpaired facial learning follows the similar constraint as in Eq. (4) but with a different input $\mathbf{P}_{y,1}^G$, and also indicates the strong correlations of face and audio:

$$\begin{aligned}\mathcal{L}_{x-y}^{un}(\mathbf{F}) &= \|D_f(E_p(\mathbf{P}_{y,1}^G), E_a(\{\mathbf{A}_{x,t}\}_1^n)) - \mathbf{F}_x^G\|_2^2 \\ &+ \frac{\lambda_f}{n-1}\sum_{t=1}^{n-1}\|(\hat{\mathbf{F}}_{x,t+1} - \hat{\mathbf{F}}_{x,t}) - (\mathbf{F}_{x,t+1}^G - \mathbf{F}_{x,t}^G)\|_2^2.\end{aligned} \tag{6}$$

Thus the overall learning objective has the form:

$$\begin{aligned}\mathcal{L}_{sum} &= \lambda_1(\mathcal{L}_x^{init} + \mathcal{L}_y^{init}) + \lambda_2(\mathcal{L}_{x-y}^{un}(\mathbf{P}) + \mathcal{L}_{x-x}^{pair}(\mathbf{P})) \\ &+ \lambda_3(\mathcal{L}_{x-y}^{un}(\mathbf{F}) + \mathcal{L}_{x-x}^{pair}(\mathbf{F})),\end{aligned} \tag{7}$$

where $\lambda_{1\sim 3}$ are the balancing weights for different losses.

Gestures2video neural rendering. In order to generate the realistic speech video for a given audio sequence, we first render synthetic human mesh images from the predicted SMPL-X model parameters. Then, the gestures2video network is employed to translate the rendered images $\hat{\mathbf{R}}$ into the final photo-realistic frames. By following previous works [12, 48, 65], we adopt a conditional-GAN architecture for our gestures2video rendering network. Specifically, to ensure the continuity of the generated frames, we use a window of size $2N_w$ with the current frame at the center of the window. The generator takes the stacked tensor $\{\hat{\mathbf{R}}_t\}_{t-N_w}^{t+N_w}$ as input and outputs a photo-realistic image $\hat{\mathbf{I}}$ of the target person. A multi-scale discriminator is designed to guarantee the quality of generated images that are optimized in an adversarial manner. The loss function for our gestures2video rendering network is defined as:

$$\begin{aligned}\mathcal{L}_{render} &= \omega_1\mathcal{L}_{VGG}(\hat{\mathbf{I}}, \mathbf{I}^G) + \omega_2\mathcal{L}_1(\hat{\mathbf{I}}, \mathbf{I}^G) \\ &+ \omega_3\mathcal{L}_{GAN} + \omega_4\mathcal{L}_{FM},\end{aligned} \tag{8}$$

where $\mathcal{L}_{VGG}$ denotes the perceptual loss that was proposed in previous image synthesis works [15, 30]. We use $\hat{\mathbf{I}}$ and

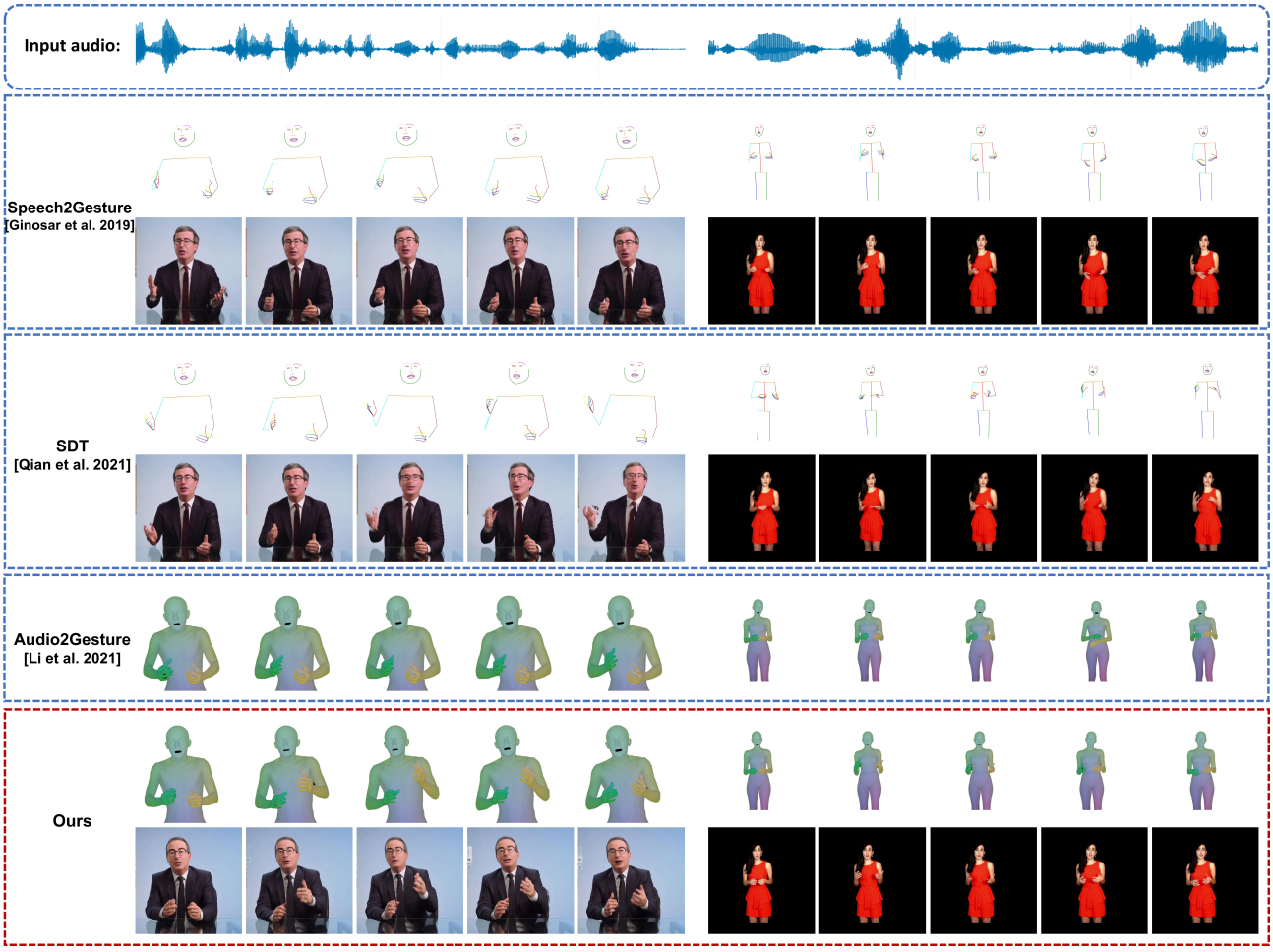

Figure 4. Comparisons of state-of-the-art models with our method. Note that Speech2Gesture [23] and SDT [49] relies on 2D skeletons as intermediate representation while Audio2gesture [36] only generates sequences of 3D models. Compared with these works, our results show sufficient gesture diversity with varying input audios. Please refer to the supplementary video for detailed qualitative comparison.

$\mathbf{I}^G$ to represent the predicted frames and original frames respectively. $\mathcal{L}_{GAN}$ represents the GAN adversarial loss and $\mathcal{L}_{FM}$ is the feature-matching loss [56] of the discriminator.

4. Experiments

Dataset. For dataset construction, we collect videos from two major sources. 1) Online videos: we use the public Oliver dataset [23] in A that is in sitting gestures, and TEDx videos in B[1] and C [2], which have speakers in standing gestures. 2) Self-captured: we capture a customized dataset with both sitting in D and standing gestures in E, forming a comprehensive benchmark. (See supplementary.)

Implementation details. All experiments are conducted on a single NVIDIA 2080-Ti GPU using Adam [31]. In audio2gestures, we use a sliding window of size $T = 128$ to extract training samples of audio and video. A total of 100 epochs are trained with a batch size of 4 and a learning rate of 0.0001. For the gestures2video network, the training takes 50 epochs with a batch size of 1 and a learning rate of 0.0001. In our experiments, the parameters used in Eqs. (3)-(8) are: $\lambda_p = 10, \lambda_f = 2, \lambda_1 = 40, \lambda_2 = 10, \lambda_3 = 40, \omega_1 = 10, \omega_2 = 50, \omega_3 = 1$, and $\omega_4 = 2$.

4.1. Comparison with State-of-the-Arts

4.1.1 Qualitative Evaluation

As shown in Fig. 4, we compare our work with three state-of-the-art methods [23, 36, 49]. In Speech2Gesture [23], only pose gestures are generated by their generator network, so we add the face part in the output layer of the generator network. SDT [49] adopts a template of gestures as the con-

[1] https://youtu.be/B99G5_T9xX4
[2] https://youtu.be/ZoLZCJHQcqU

Table 2. Quantitative comparisons to state-of-the-art methods [23, 36, 49]. Best values are highlighted in bold.

Dataset	Method	Offset/confidence	LPIPS	CPBD	FVD	Diversity	Multimodality
Online videos	Speech2Gesture [23]	-3/1.751	0.267	0.520	636.7	6.430	-
	SDT [49]	-3/1.923	0.253	0.511	544.0	8.810	7.796
	Audio2Gesture [36]	-	-	-	-	13.647	11.747
	Ours	-2/**2.328**	**0.156**	**0.569**	**387.3**	**16.915**	**15.931**
Self-captured videos	Speech2Gesture [23]	1/0.570	0.196	0.492	310.1	8.913	-
	SDT [49]	-2/1.275	0.187	0.502	302.4	8.513	8.159
	Audio2Gesture [36]	-	-	-	-	11.318	12.553
	Ours	1/**2.029**	**0.135**	**0.526**	**246.5**	**13.447**	**16.472**

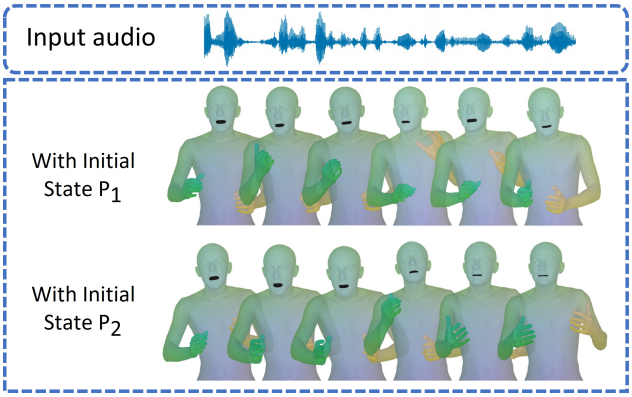

Figure 5. Illustrations of our one-to-many pose generation. With the same audio but different initial poses, our disentangled representation learning can generate different diverse sequences.

ditional input, which can determine the general appearance of the generated gesture sequence. Both Speech2Gesture and SDT are keypoint-based methods. By using 3D dataset, Audio2Gesture [36] predicts the gestures defined on 3D skeleton. In contrast, with Disentangled Recurrent Representation (DR^2) Learning, our method generates diversified and photo-realistic talking videos with gestures and face motions. Please refer to our supplementary video for a detailed comparison.

4.1.2 Quantitative Evaluation

Following prevailing works, we adopt the lip sync, image quality, and gesture diversity metrics for comparisons. Tab. 2 shows the comparison with state-of-the-art methods, indicating the superior performance of our models.

Lip-sync metric: We evaluate the synchronization of lip motion with input audio by SyncNet [19], which calculates the Audio-Visual (AV) Offset and Confidence scores to determine the lip-sync error. Under this challenging lip-syncing with limited data, our results are better than the baseline, which is attributed to our approach of separating the learning of pose and face for high stability.

Image quality metric: The image quality is evaluated by learned perceptual image patch similarity (LPIPS) and cumulative probability blur detection (CPBD). Our results demonstrate superior performance in terms of image clarity and sharpness.

Temporal-level metric: We apply Fréchet Video Distance (FVD) to evaluate the realism of results at temporal level. Compared to baselines, FVD not only measures the quality of generated videos but also indicates that our video gesture distribution is closer to that of real videos.

Diversity: We adopt the gesture diversity metric [36] to evaluate how many different poses/motions have been generated within a long sequence. For a fair comparison, we project the human model joints into 2D space and select the corresponding keypoints for measurement. From the diversity result, our training significantly preserves the effectiveness of gestures compared to the baselines. This is due to our disentangled training approach and the incorporation of elements like noise to expand training possibilities.

Multi-modal diversity: We measure the multi-modal diversity [36] by generating motion sequences N times given the same audio and then calculating the average L_1 distance of N motion sequences. As there is a one-to-many mapping relationship between audio and gesture, our method also achieves robust results in terms of multi-modality.

4.2. Ablations and Performance Analysis

Disentangled representation learning. To evaluate the effectiveness of our disentangled representation learning, we conduct ablation studies as in Tab. 3. The baseline model in the first line indicates the conventional one-to-one learning with paired audio and pose sequences, which shows less sequence diversity during speech and cannot generate different outputs with the same input, *i.e.*, '-' in the multi-modal diversity. In the second row, as we only use the one-shot

Table 3. Ablation for disentangled learning on Oliver data. Each line adds a new component compared to its previous line.

Method	Diversity ↑	Multimodality ↑
Baseline	8.753	-
+Initial State	8.036	9.943
+Disentangled Training	21.749	23.740
+Random Noise	**23.055**	**24.898**

frame as initial input, the long sequence diversity shows a slight performance drop. However, the employment of the initial state enables the generation of one-to-many mappings. Moreover, it provides the continuity of adjacent generated clips. As the most crucial part of our disentangled representations, adding the disentangled training module in the third row greatly enhances the diversity of generated long sequences and multiple modalities. Besides, the results in the last row show that adding random noise can further benefit the generalization of various inference scenarios.

One-to-many generation. Existing models usually rely on the paired "one-to-one" mapping for learning pose sequences, while the disentangled representation learning enables our network with the "one-to-many" generation ability. Fig. 5 visualizes the generated sequences with two different initial poses $\mathbf{P}_1$ and $\mathbf{P}_2$. It shows that our method can generate different high-quality sequences with the sample input audio by modifying the initial input with any possible random state.

Motion-audio correlation. Beyond the diversity exhibited by our method, a high-quality synthesized result should also demonstrate motion-audio correlations that are close to ground-truth cases. Therefore, landmark velocity difference [57, 68] is adopted to evaluate the speech-gestures correlation. We designed the ablation study with three different settings: random unpaired sequence, **w/o** motion loss in $\mathcal{L}^{un}_{x-y}(\mathbf{P})$, and **w/** motion loss (ours). The results in Tab. 4 verify the synchronization of our co-speech gestures.

4.3. User Studies

To gauge the quality of generated videos from a human-centric standpoint, we carried out a user study involving 20 volunteers. The evaluation criteria consisted of four key aspects: Audio-Lip Synchronization, Photo-realistic Image Quality, Gesture Diversity, and Overall Preference. Each participant was instructed to evaluate every video four times, once for each criterion. We employed a scoring system ranging from 0 (Very Poor) to 4 (Excellent). Each volunteer reviewed a set of 12 video clips with a length of 10-20 seconds. Each video contains two baselines and our method. We calculated the average evaluation scores for each criterion and for each method. The summarized results, presented in Fig. 6, reveal that our approach outperforms state-of-the-art methods from the human evaluators.

Table 4. Motion-audio correlation study on Oliver data. Each testing included 25 sequences and we recorded the average score and the best score (in parentheses).

Method	Landmark velocity difference ↓
Random sequence	21.941 (20.442)
w/o motion loss	19.802 (17.987)
w/ motion loss	**18.401 (16.785)**

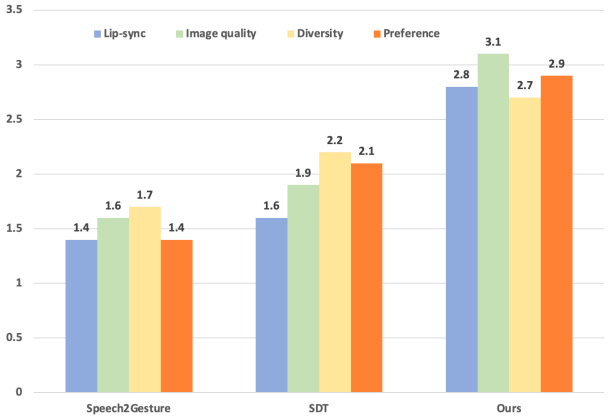

Figure 6. User Study. We calculated the average scores (ranging from 0 to 4) given by users across four evaluation metrics for two baselines and our method. The results show that our method outperforms the baselines.

5. Conclusion and Future Work

In this paper, we make the disentangled audio-pose latent space assumptions for training a full-body talking video with only short-term video dataset of only 2 minutes. Based on this assumption, we develop the disentangled training module and infinite inference module for generating long-term diverse co-speech gestures. In our future work, we would like to model the explicit mutual effect of audio and pose latent embedding and explore the diverse generation with more challenging "in the wild" data. For example, modeling various speech scenarios by training with dozens of short video clips of only a few seconds.

Acknowledgments

Zhang and Guo were partially supported by National Science Foundation (OAC-2007661) and Cisco Research. The opinions expressed are solely those of the authors, and do not necessarily represent those of the funding agencies.

References

[1] 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 *ECCV*, 2020. 2

[2] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *3DV*, 2019. 2

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

[4] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM TOG*, 2022. 3

[5] Tenglong Ao, Zeyi Zhang, and Libin Liu. Gesturediffuclip: Gesture diffusion model with clip latents. *ACM TOG*, 2023. 3

[6] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *CVPR*, 2018. 3

[7] Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 2

[8] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *VR*, 2021. 2

[9] Elif Bozkurt, Shahriar Asta, Serkan Özkul, Yücel Yemez, and Engin Erzin. Multimodal analysis of speech prosody and upper body gestures using hidden semi-markov models. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013. 2

[10] Justine Cassell. Embodied conversational interface agents. *Communications of the ACM*, 2000. 1, 2

[11] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 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*, 1994. 2

[12] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *ICCV*, 2019. 3, 5

[13] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *ECCV*, 2020. 3

[14] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *CVPR*, 2019. 1, 3

[15] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, 2017. 5

[16] Chung-Cheng Chiu and Stacy Marsella. How to train your avatar: A data driven approach to gesture generation. In *International Workshop on Intelligent Virtual Agents*, 2011. 2

[17] Chung-Cheng Chiu and Stacy Marsella. Gesture generation with low-dimensional embeddings. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 2014. 2

[18] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? In *BMVC*, 2017. 3

[19] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *ACCV*, 2016. 7

[20] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *CVPR*, 2019. 3

[21] Ylva Ferstl, Michael Neff, and Rachel McDonnell. Adversarial gesture generation with realistic gesture phasing. *Computers & Graphics*, 2020. 2

[22] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F. Troje, and Marc-André Carbonneau. Zeroeggs: Zero-shot example-based gesture generation from speech. *Computer Graphics Forum*, 2023. 3

[23] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *CVPR*, 2019. 1, 2, 3, 6, 7

[24] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *ICCV*, 2021. 1, 3

[25] Ikhsanul Habibie, Mohamed Elgharib, Kripasindhu Sarkar, Ahsan Abdullah, Simbarashe Nyatsanga, Michael Neff, and Christian Theobalt. A motion matching-based framework for controllable gesture synthesis from speech. In *ACM SIGGRAPH 2022 Conference Proceedings*, 2022. 2

[26] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, 2021. 2

[27] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014. 5

[28] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. Evaluation of speech-to-gesture generation using bi-directional lstm network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, 2018. 2

[29] Carlos T Ishi, Daichi Machiyashiki, Ryusuke Mikata, and Hiroshi Ishiguro. A speech-driven hand gesture generation method and evaluation in android robots. *IEEE Robotics and Automation Letters*, 2018. 2

[30] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 5

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[32] Taras Kucherenko, Dai Hasegawa, Naoshi Kaneko, Gustav Eje Henter, and Hedvig Kjellström. Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation. *International Journal of Human–Computer Interaction*, 2021. 2

[33] Taras Kucherenko, Patrik Jonell, Sanne Van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 international conference on multimodal interaction*, 2020. 3

[34] Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. Gesture controllers. In *ACM SIGGRAPH 2010 papers*. 2010. 2

[35] Sergey Levine, Christian Theobalt, and Vladlen Koltun. Real-time prosody-driven synthesis of body language. In *ACM SIGGRAPH Asia 2009 papers*. 2009. 2

[36] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *CVPR*, 2021. 1, 2, 6, 7

[37] Yuanzhi Liang, Qianyu Feng, Linchao Zhu, Li Hu, Pan Pan, and Yi Yang. Seeg: Semantic energized co-speech gesture generation. In *CVPR*, 2022. 3

[38] Miao Liao, Sibo Zhang, Peng Wang, Hao Zhu, Xinxin Zuo, and Ruigang Yang. Speech2video synthesis with 3d skeleton regularization and expressive body poses. In *ACCV*, 2020. 1, 3

[39] Haiyang Liu, Naoya Iwamoto, Zihao Zhu, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Disco: Disentangled implicit content and rhythm learning for diverse co-speech gestures synthesis. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 3

[40] 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 *ECCV*, 2022. 3

[41] Xian Liu, Qianyi Wu, Hang Zhou, Yuanqi Du, Wayne Wu, Dahua Lin, and Ziwei Liu. Audio-driven co-speech gesture video generation. In *Advances in Neural Information Processing Systems*. 3

[42] 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 *CVPR*, 2022. 1, 2, 3

[43] Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live speech portraits: Real-time photorealistic talking-head animation. *ACM TOG*, 2021. 3

[44] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 2013. 1, 2

[45] Victor Ng-Thow-Hing, Pengcheng Luo, and Sandra Okita. Synchronized gesture and speech production for humanoid robots. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010. 1, 2

[46] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 2, 3, 5

[47] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM MM*, 2020. 3

[48] Sergey Prokudin, Michael J Black, and Javier Romero. Smplpix: Neural avatars from 3d human models. In *WACV*, 2021. 5

[49] Shenhan Qian, Zhi Tu, Yihao Zhi, Wen Liu, and Shenghua Gao. Speech drives templates: Co-speech gesture synthesis with learned templates. In *ICCV*, 2021. 1, 2, 3, 6, 7

[50] Najmeh Sadoughi, Yang Liu, and Carlos Busso. Speech-driven animation constrained by appropriate discourse functions. In *Proceedings of the 16th International Conference on Multimodal Interaction*, 2014. 2

[51] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. Audio to body dynamics. In *CVPR*, 2018. 2

[52] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM TOG*, 2017. 3

[53] Kenta Takeuchi, Dai Hasegawa, Shinichi Shirakawa, Naoshi Kaneko, Hiroshi Sakuta, and Kazuhiko Sumi. Speech-to-gesture generation: A challenge in deep learning approach with bi-directional lstm. In *Proceedings of the 5th International Conference on Human Agent Interaction*, 2017. 2

[54] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *ECCV*, 2020. 1, 3

[55] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3

[56] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 6

[57] Jing Xu, Wei Zhang, Yalong Bai, Qibin Sun, and Tao Mei. Freeform body motion generation from speech. *arXiv preprint arXiv:2203.02291*, 2022. 8

[58] 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. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, 2023. 3

[59] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, and Haolin Zhuang. Qpgesture: Quantization-based and phase-guided motion matching for natural speech-driven gesture generation. In *CVPR*, 2023. 3

[60] Sheng Ye, Yu-Hui Wen, Yanan Sun, Ying He, Ziyang Zhang, Yaoyuan Wang, Weihua He, and Yong-Jin Liu. Audio-driven stylized gesture generation with flow-based model. In *ECCV*, 2022. 2

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

[62] 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 TOG*, 2020. 3

[63] 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 *ICRA*, 2019. 2

[64] Chenxu Zhang, Saifeng Ni, Zhipeng Fan, Hongbo Li, Ming Zeng, Madhukar Budagavi, and Xiaohu Guo. 3d talking face with personalized pose dynamics. *IEEE TVCG*, 2021. 1, 3

[65] Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. Facial: Synthesizing dynamic talking face with implicit attribute learning. In *ICCV*, 2021. 3, 5

[66] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*, 2019. 1, 3

[67] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *CVPR*, 2021. 3

[68] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM TOG*, 2020. 1, 3, 8

[69] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *CVPR*, 2023. 3

[70] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *arXiv preprint arXiv:2307.10894*, 2023. 3