Generating Holistic 3D Human Motion from Speech

Hongwei Yi[1*] Hualin Liang[2*] Yifei Liu[2*] Qiong Cao[3†]
Yandong Wen[1] Timo Bolkart[1] Dacheng Tao[3] Michael J. Black[1†]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]South China University of Technology [3]JD Explore Academy
{hongwei.yi, yandong.wen, timo.bolkart, black}@tuebingen.mpg.de
{hualinliang3, yifei9697, mathqiong2012, dacheng.tao}@gmail.com

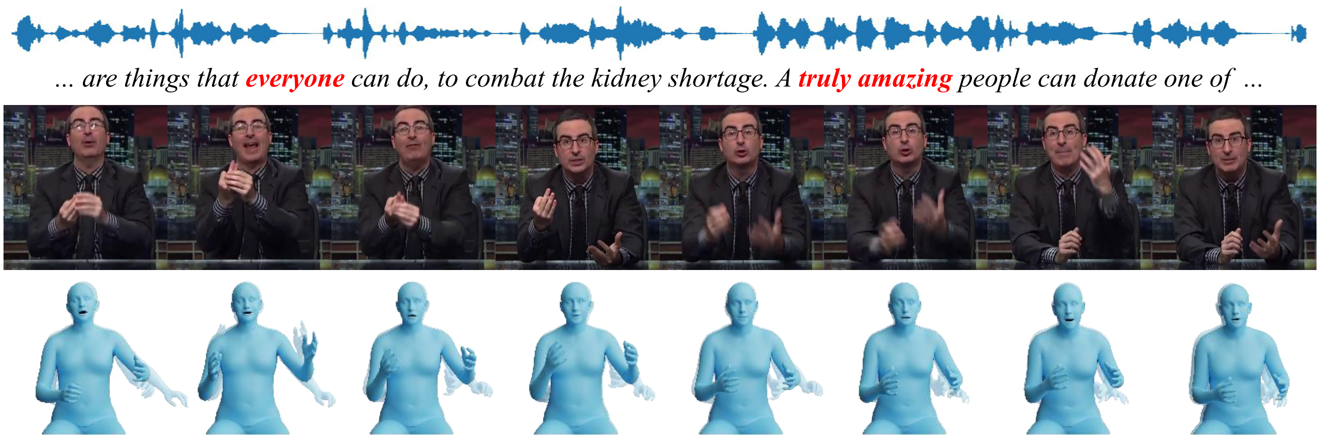

Figure 1. **Speech-to-motion translation example.** Given a speech signal as input, our approach generates realistic, coherent, and diverse holistic body motions; that is, the body motion together with facial expressions and hand gestures. From top to bottom: the input audio, the corresponding transcript, video frames, and the generated motions. Note that the audio is the only input to our approach, while the transcript and video frames are just shown for reference.

Abstract

This work addresses the problem of generating 3D holistic body motions from human speech. Given a speech recording, we synthesize sequences of 3D body poses, hand gestures, and facial expressions that are realistic and diverse. To achieve this, we first build a high-quality dataset of 3D holistic body meshes with synchronous speech. We then define a novel speech-to-motion generation framework in which the face, body, and hands are modeled separately. The separated modeling stems from the fact that face articulation strongly correlates with human speech, while body poses and hand gestures are less correlated. Specifically, we employ an autoencoder for face motions, and a compositional vector-quantized variational autoencoder (VQ-VAE) for the body and hand motions. The compositional VQ-VAE is key to generating diverse results. Additionally, we propose a cross-conditional autoregressive model that generates body poses and hand gestures, leading to coherent and realistic motions. Extensive experiments and user studies demonstrate that our proposed approach achieves state-of-the-art performance both qualitatively and quantitatively. Our dataset and code are released for research purposes at `https://talkshow.is.tue.mpg.de/`.

1. Introduction

From linguistics and psychology we know that humans use body language to convey emotion and use gestures in communication [22, 28]. Motion cues such as facial expression, body posture and hand movement all play a role. For instance, people may change their gestures when shifting to a new topic [52], or wave their hands when greeting an audience. Recent methods have shown rapid progress on modeling the translation from human speech to body motion, and can be roughly divided into rule-based [38] and learning-based [20, 21, 23, 32, 33, 55] methods. Typically, the body motion in these methods is represented as the motion of a 3D mesh of the face/upper-body [5, 15, 26, 43, 44], or 2D/3D landmarks of the face with 2D/3D joints of the

*Equal Contribution.
†Joint Corresponding Authors.

hands and body [21, 23, 55]. However, this is not sufficient to understand human behavior. Humans communicate with their bodies, hands and facial expressions together. Capturing such coordinated activities as well as the full 3D surface in tune with speech is critical for virtual agents to behave realistically and interact with listeners meaningfully.

In this work, we focus on generating the expressive 3D motion of person, including their body, hand gestures, and facial expressions, from speech alone; see Fig. 1. To do this, we must learn a cross-modal mapping between audio and 3D holistic body motion, which is very challenging in practice for several reasons. First, datasets of 3D holistic body meshes and synchronous speech recordings are scarce. Acquiring them in the lab is expensive and doing so in the wild has not been possible. Second, real humans often vary in shape, and their faces and hands are highly deformable. It is not trivial to generate both realistic and stable results of 3D holistic body meshes efficiently. Lastly, as different body parts correlate differently with speech signals, it is difficult to model the cross-modal mapping and generate realistic and diverse holistic body motions.

We address the above challenges and learn to model the conversational dynamics in a data-driven way. Firstly, to overcome the issue of data scarcity, we present a new set of 3D holistic body mesh annotations with synchronous audio from in-the-wild videos. This dataset was previously used for learning 2D/3D gesture modeling with 2D body keypoint annotations [21] and 3D keypoint annotations of the holistic body [23] by applying existing models separately. Apart from facilitating speech and motion modeling, our dataset can also support broad research topics like realistic digital human rendering. Then, to support our data-driven approach to modeling speech-to-motion translation, an accurate holistic body mesh is needed. Existing methods have focused on capturing either the body shape and pose isolated from the hands and face [8, 17, 25, 34, 47, 57, 58, 61], or the different parts together, which often produces unrealistic or unstable results, especially when applied to video sequences [18, 40, 62]. To solve this, we present SHOW, which stands for "Synchronous Holistic Optimization in the Wild". Specifically, SHOW adapts SMPLify-X [40] to the videos of talking persons, and further improves it in terms of stability, accuracy, and efficiency through careful design choices. Figure 2 shows example reconstruction results.

Lastly, we investigate the translation from audio to 3D holistic body motion represented as a 3D mesh (Fig. 1). We propose TalkSHOW, the first approach to autoregressively synthesize realistic and diverse 3D body motions, hand gestures and facial expression of a talking person from speech. Motivated by the fact that the face (i.e. mouth region) is strongly correlated with the audio signal, while the body and hands are less correlated, or even uncorrelated, TalkSHOW designs separate motion generators for different parts and gives each part full play. For the face part, to model the highly correlated nature of phoneme-to-lip motion, we design a simple encoder-decoder based face generator that encodes rich phoneme information by incorporating the pre-trained wav2vec 2.0 [6]. On the other hand, to predict the non-deterministic body and hand motions, we devise a novel VQ-VAE [50] based framework to learn a compositional quantized space of motion, which efficiently captures a diverse range of motions. With the learned discrete representation, we further propose a novel autoregressive model to predict a multinomial distribution of future motion, cross-conditioned between existing motions. From this, a wide range of motion modes representing coherent poses can be sampled, leading to realistic looking motion generation.

We quantitatively evaluate the realism and diversity of our synthesized motion compared to ground truth and baseline methods and ablations. To further corroborate our qualitative results, we evaluate our approach through an extensive user study. Both quantitative and qualitative studies demonstrate the state-of-the-art quality of our speech-synthesized full expressive 3D character animations.

2. Related work

2.1. Holistic Body Reconstruction

Recent work addresses the problem of 3D holistic body mesh recovery [12, 24, 40, 54, 62]. SMPLify-X [40] fits the parametric and expressive SMPL-X model [40] to 2D keypoints obtained by off-the-shelf detectors (e.g. OpenPose [9]). PIXIE [18] directly regresses SMPL-X parameters using moderators that estimate the confidence of part-specific features. These features are fused and fed to independent regressors. PyMAF-X [62] improves the body and hand estimation with spatial alignment attention. In this work, we adapt the optimization-based SMPLify-X to videos of talking persons, and improve the stability and accuracy with several good engineering practices in terms of initialization, data term design, and regularization.

2.2. Speech-to-Motion Datasets

The existing speech-to-motion datasets can be roughly categorized as in-house and in-the-wild. The annotations of in-house datasets [13, 16, 20, 48, 53] are accurate but are limited in scale since the multi-camera systems used for data capture are expensive and labor intensive. Moreover, these datasets only provide annotations of the head [13, 16, 53] or body [20, 48], and thus do not support whole-body generation. To learn richer and more diverse speaking styles and emotions, [59, 60] propose to use in-the-wild videos. The annotations are pseudo ground truth (p-GT) given by advanced reconstruction approaches, *e.g.* [9]. However, these released datasets use either 2D keypoints or 3D keypoints with 3D head mesh to represent the body. This disconnected repre-

Dataset	Head	Hand	Body	Holistic Body Connection	In-the-wild	Length	Annotations
Multiface [53]	3D mesh	✘	✘	✘	✘	-	multi-camera
BIWI [16]	3D mesh	✘	✘	✘	✘	-	3D-scanner
VOCASET [13]	3D mesh	✘	✘	✘	✘	-	4D-scan
Takeuchi et.al [48]	✘	✘	3D keypoint	✘	✘	5h	MoCap
Trinity [20]	✘	✘	3D keypoint	✘	✘	4h	MoCap
Yoon et.al [59, 60]	✘	✘	3D keypoint	✘	✔	52h	p-GT
Speech2Gesture [21]	✘	2D keypoint	2D keypoint	✘	✔	144h	p-GT
Habibie et.al [23]	3D mesh	3D keypoint	3D keypoint	✘	✔	33h	p-GT
Ours	3D mesh	3D mesh	3D mesh	✔	✔	27h	p-GT

Table 1. Comparison of different speech-to-motion datasets.

sentation limits the possible applications of the generated talking motions. In contrast to the aforementioned work, our dataset, reconstructed by SHOW, consists of holistic body meshes and synchronized speech, covering a wide range of body poses, hand gestures, and facial expressions. More details can be found in Table 1.

2.3. Holistic Body Motion Generation from Speech

Holistic body motion generation from speech consists of three body parts motion generation, i.e., faces, hands, and bodies. Existing 3D talking face generation methods [13, 15, 43, 63] rely heavily on 4D face scan datasets for training [13, 16, 43]. There are many attempts to perform body motion generation, and these can be divided into rule-based and learning-based methods. Rule-based methods [10, 29, 33, 41] map the input speech to pre-collected body motion “units” with manually designed rules. They are explainable and controllable but it is expensive to create complex, realistic, motion patterns. Learning-based body motion generation approaches [1, 7, 21, 30, 31, 36, 59] have advanced significantly in part due to publicly released synchronous speech and body motion datasets [21, 23, 35, 48, 59, 60]. However, they only consider parts of the human body rather than the holistic body. Most related to our work, Habibie et al. [23] propose to generate 3D facial meshes and 3D keypoints of the body and hands from speech, but the generated faces, bodies and hands are disjoint. Also, these methods are deterministic, can not generate diverse motions when given the same speech recording. There are a few attempts to incorporate the diversity into motion generation using GANs [2, 37, 60], VAEs [35, 42, 55], VQ-VAEs [5, 56], or normalising-flows [3]. Nevertheless, the diversity of motions produced by these methods is inadequate.

In contrast, TalkSHOW generates holistic body motions and models different body parts separately according to their natures: the face part is more correlated to the speech signal than body parts. TalkSHOW develops a simple deterministic encoder-decoder structure for mapping acoustic signals to facial expressions. TalkSHOW adopts two VQ-VAEs to generate more diverse body and hand motions. This novel design allows the learned quantized space to be compositional and more expressive for conversational gestures. Compared with previous VQ-VAE-based methods [46, 56], we design a cross-conditional autoregressive model to generate different body-part motions, which are more fluid and natural.

3. Dataset

In this section, we introduce a high-quality audiovisual dataset, which consists of expressive 3D body meshes at 30fps, and their synchronized audio at a 22K sample rate. The 3D body meshes are reconstructed from in-the-wild monocular videos and are used as our pseudo ground truth (p-GT) in speech-to-motion generation. We provide detailed descriptions of this dataset in Sec. 3.1 and highlight several good practices for obtaining more accurate p-GT from videos in Sec. 3.2. Our experiments show that this dataset is effective for training speech-to-motion models.

3.1. Dataset Description

The dataset is built from the in-the-wild talking videos of different people with various speaking styles. We use the same video sources from [21] for straightforward comparisons with the previous work. To facilitate the subsequent 3D body reconstruction, we manually filter out videos if they are in any following cases: (i) low resolution (<720p), (ii) occluded hand(s), or (iii) invalid download link. The filtering leads to a high-quality dataset of 26.9 hours from 4 speakers. For the mini-batch processing, the raw videos are cropped into short clips (<10 seconds). Direct comparisons to the existing datasets can be found in Table 1.

Expressive 3D whole-body meshes are reconstructed from these videos and used as the p-GT. Specifically, the 3D holistic body meshes consist of face, hands, and bodies in a connected way, which is achieved by adopting a well-designed 3D topology from SMPL-X [40]. As a result, we represent the p-GT of the dataset as SMPL-X parameters. Given a video clip of T frames, the p-GT comprises parameters of a shared body shape $\beta \in \mathbb{R}^{300}$, poses $\{\theta_t | \theta_t \in \mathbb{R}^{156}\}_{t=1}^{T}$, a shared camera pose $\theta^c \in \mathbb{R}^3$ and translation $\epsilon \in \mathbb{R}^3$, and facial expressions $\{\psi_t | \psi_t \in \mathbb{R}^{100}\}_{t=1}^{T}$.

Figure 2. The 3D holistic body reconstruction results from SMPLify-X, PIXIE, PyMAF-X, and ours. Compared to other methods, ours produces more accurate and stable results with details.

Here the pose θ_t includes jaw pose $\theta_t^{jaw} \in \mathbb{R}^3$, body pose $\theta_t^b \in \mathbb{R}^{63}$, and hand pose $\theta_t^h \in \mathbb{R}^{90}$.

We note that this dataset can not only be used in speech and motion modeling, but also supports broad research topics like realistic digital human rendering and learning-based holistic body recovery from videos, etc.

3.2. Good Practices for Improving p-GT

In this section, we present SHOW, which adapts SMPLify-X [40] to the videos of talking persons with several good practices, to improve the stability, accuracy, and efficiency in 3D whole-body reconstruction. In the following, we briefly summarize our efforts for improving the p-GT. See more details in the supplemental material.

Initialization. A good initialization can significantly accelerate and stabilize the SMPLify-X optimization. We apply several advanced regression-based approaches to the videos, and use the resulting predictions as the initial parameters of SMPLify-X. Specifically, PIXIE [18], PyMAF-X [62], and DECA [19] are used to initialize θ^b, θ^h, and θ^f, respectively. The camera is assumed to be static, and its parameters θ^c and ϵ are estimated by PIXIE [18] as well.

Data Term. The joint re-projection loss is the most important data objective function in SMPLify-X, as it optimizes the difference between joints extracted from the SMPL-X model, projected into the image, with joints predicted with OpenPose [9]. Here we extend the data term by incorporating body silhouettes from DeepLab V3, facial landmarks from MediaPipe [27], and facial shapes from MICA [65]. Further, we use a photometric loss between the rendered faces and the input image to better capture facial details.

Regularization. Different regularization terms in SMPLify-X prevent the reconstruction of unrealistic bodies. To derive more reasonable regularizations, we explicitly take information about the video into account and make the following assumptions. First, the speaker in each video clip remains the same. This is further verified by a face recognition pipeline using the ArcFace model [14]. So we can use consistent shape parameters β to represent the holistic body shape. Second, the holistic body pose, facial expression, and environmental lighting in video clips change smoothly over time. This temporal smoothness assumption has proven useful in many previous approaches [58, 65], and we observe similar improvements in our experiments. Third, the person's surface does self-penetrate, which should be self-evident in the real world.

Overall, as shown in Figure 2, the p-GT can be significantly improved by incorporating the aforementioned practices. See more results in the supplemental video.

4. Method

Given a speech recording, our goal is to generate conversational body poses, hand gestures as well as facial expressions that match the speech in a plausible way. Motivated by the fact that the face motion is highly correlated to the speech signal, while the body and hand parts are less correlated, we propose TalkSHOW, a novel framework that can model speech and different human parts separately. In the following, we present an encoder-decoder based face generator in Sec. 4.2, and a body and hand generator in Sec. 4.3.

4.1. Preliminary

Let $M = \{m_t\}_{t=1}^T$ be a p-GT holistic motion sequence (i.e., a temporal sequence of the poses m_i) provided in Sec. 3. We denote the motion of the face, body and hands as M^f, M^b and M^h respectively, see more details in supplemental material.

4.2. Face Generator

Given a raw audio signal $A_{1:T}$, our face generator G_F aims to generate expressive facial motions $\widehat{M}^f_{1:T} = (\widehat{m}^f_1, \ldots, \widehat{m}^f_T) \in \mathbb{R}^{103\times T}$ close to $M^f_{1:T} \in \mathbb{R}^{103\times T}$.

Figure 3 (A) illustrates our idea. In order to produce synchronized mouth motions [15], we leverage a pretrained speech model, wav2vec 2.0 [6]. Specifically, the encoder consists of an audio feature extractor and a transformer encoder [51], leading to a 768-dimensional speech representation. A linear projection layer is added on top of the encoder to reduce the dimension to 256. Our decoder comprises six layers of temporal convolutional networks (TCNs) followed by a fully-connected layer. We train the encoder and decoder with an Mean Square Error (MSE) loss.

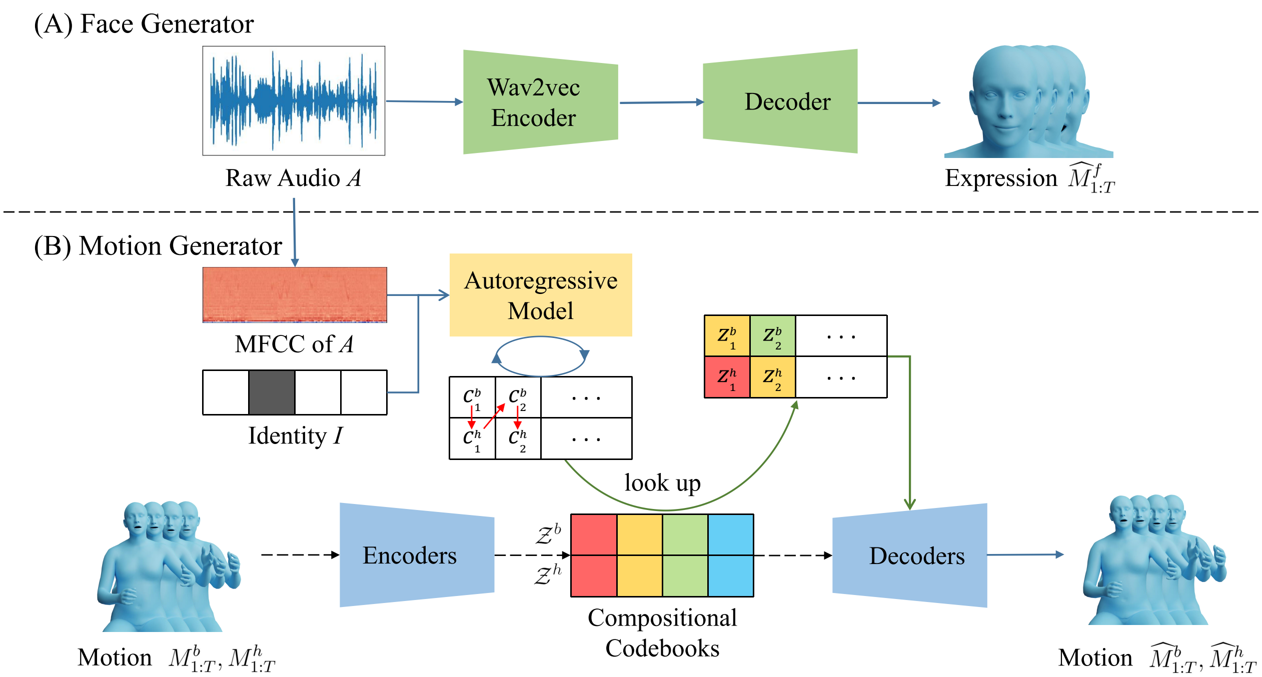

Figure 3. Overview of the proposed TalkSHOW. We employ a simple encoder-decoder model for face motions, and a novel framework for body and hand motions. Specifically, this framework first learns VQ-VAEs on each piece separately to obtain a compositional quantized space. Then, we autoregressively predict a distribution of the body/hand motion at a future timestep. Our predictor is designed to be cross-conditioned between the body and hand motions to keep the synchronization of the holistic body. Lastly, we obtain the future body/hand motions by decoding a sampled codebook index sampled from the distribution. The blue, black, and green lines indicate processes involved at training and inference, training, and inference stages, respectively. Best viewed in color.

4.3. Body and Hand Generator

Given an audio input, we aim to generate a temporal sequence of realistic and diverse motions for the body and hands, *i.e.* $\widehat{M}^b_{1:T} = (\widehat{m}^b_1, \ldots, \widehat{m}^b_T) \in \mathbb{R}^{63\times T}$ and $\widehat{M}^h_{1:T} = (\widehat{m}^h_1, \ldots, \widehat{m}^h_T) \in \mathbb{R}^{90\times T}$, respectively. Figure 3 (B) illustrates our idea. Instead of learning a direct mapping from audio to motion, we leverage the recent advances of VQ-VAE [50] to learn a multi-mode distribution space for body and hand motions. Specifically, we first encode and quantize the body and hand motions into two finite codebooks, from which we can sample a wide range of plausible body and hand combinations. Then, we introduce a novel cross-conditional autoregressive model over the learned codebooks, which allows us to predict diverse body and hand motions. Our predictor is designed to be cross-conditioned between the body and hand to keep the synchronization of the holistic body. Lastly, we obtain the future body/hand motions by decoding codebook indices sampled from the distribution.

Representation. We use 64-dimensional MFCC features [45] as the audio representation for body and hand generation, i.e., $A_{1:T} = (a_1, \ldots, a_T) \in \mathbb{R}^{64\times T}$. Since body and hand gestures are more correlated to the rhythm and beat instead of phonemes, low-dimensional MFCC features are sufficient to produce plausible gestures from audio. Besides, considering that speakers often present different motion styles, we also leverage the modality of speaker identity I to differentiate those styles. We represent this as a one-hot vector $I \in \{0, 1\}^{N_I}$, where N_I is the number of speakers.

Compositional Quantized Motion Codebooks. The vanilla VQ-VAE learns a discrete codebook $\mathcal{Z} = \{z_i\}_{i=1}^{|\mathcal{Z}|}$ consisting of multiple vectors $z_i \in \mathbb{R}^{d_z}$ to quantize the latent space of input. To further expand the range that the learned codebook can represent, we divide the motions into compositional pieces, i.e., body and hands, and learn VQ-VAEs on each piece separately. By doing this, the body and hand movements are encoded and quantized into two separate finite codebooks $\mathcal{Z}^b = \{z^b_i\}_{i=1}^{|\mathcal{Z}_b|}$ and $\mathcal{Z}^h = \{z^h_j\}_{j=1}^{|\mathcal{Z}_h|}$, where $z^b_i, z^h_i \in \mathbb{R}^{d_z}$ with lengths $|\mathcal{Z}^b|$ and $|\mathcal{Z}^h|$ respectively, from which we can combine $|\mathcal{Z}^b| \times |\mathcal{Z}^h|$ different body-hands pose code pairs (z^b_i, z^h_j) to expand motion diversity. In this scheme, given an input of the sequence of body and hand motions $M^b_{1:T} \in \mathbb{R}^{63\times T}$ and $M^h_{1:T} \in \mathbb{R}^{90\times T}$, we first encode them into the feature sequence $E^b_{1:\tau} = (e^b_1, \ldots, e^b_\tau) \in \mathbb{R}^{64\times\tau}$ and $E^h_{1:\tau} = (e^h_1, \ldots, e^h_\tau) \in \mathbb{R}^{64\times\tau}$, where $\tau = \frac{T}{w}$ and w is the temporal window size. Then, we quantize the embedding by mapping it into the nearest code in the corresponding codebook:

$$
\begin{aligned}
z_t^b &= \arg\min_{z_k^b \in \mathcal{Z}^b} \|e_t^b - z_k^b\| \in \mathbb{R}^{64}, \\
z_t^h &= \arg\min_{z_k^h \in \mathcal{Z}^h} \|e_t^h - z_k^h\| \in \mathbb{R}^{64}.
\end{aligned} \tag{1}
$$

Finally, the quantized features $Z_{1:\tau}^b = (z_1^b, \ldots, z_\tau^b) \in \mathbb{R}^{64\times\tau}$ and $Z_{1:\tau}^h = (z_1^h, \ldots, z_\tau^h) \in \mathbb{R}^{64\times\tau}$ are fed into the decoder for the synthesis.

We train the encoder and decoder simultaneously with the codebook by the following loss function:

$$
\begin{aligned}
\mathcal{L}_{VQ} = \mathcal{L}_{rec}(M_{1:T}, \widehat{M}_{1:T}) + \|\mathrm{sg}[E_{1:T}] - Z_{1:T}\| \\
+ \beta \|E_{1:T} - \mathrm{sg}[Z_{1:T}]\|,
\end{aligned} \tag{2}
$$

where $\mathcal{L}_{rec}$ is an MSE reconstruction loss, sg is a stop gradient operation [11] that is used to calculate codebooks loss, and the third part is a "commitment" loss with a trade-off β.

Cross-Conditional Autoregressive Modeling. After we learn the compositional quantized codebooks, any body and hand motions can be represented as a sequence of codebook vectors via the encoder and quantization. Following the common paradigm, we use the representation of a sequence of corresponding codebook indices, in the form of one-hot vectors, of the nearest codebook entry per element, which is denoted as $C_{1:\tau}^b = (c_1^b, \ldots, c_\tau^b) \in \mathbb{R}^{|\mathcal{Z}^b|\times\tau}$ and $C_{1:\tau}^h = (c_1^h, \ldots, c_\tau^h) \in \mathbb{R}^{|\mathcal{Z}^h|\times\tau}$.

Now, with the quantized motion representation, we design a temporal autoregressive model over it to predict the distribution of possible next motions, given the input audio embedding A and existing motions. Besides, we enable the modality input of identity I to distinguish different gesture styles. Because we model the body and hands independently, to keep the consistency of the holistic body and thus predict realistic gestures, we exploit the mutual information and design our model to be cross-conditioned between the body and hand motions. Specifically, following Bayes' Rule, we model the joint probability of $C_{1:\tau}^b$ and $C_{1:\tau}^h$ as follows:

$$
\begin{aligned}
p(C_{1:\tau}^b, C_{1:\tau}^h \mid A_{1:\tau}, I) = \prod_{t=1}^{\tau} p\left(c_t^b \mid c_{<t}^b, c_{<t}^h, a_{\leq t}, I\right) \\
p\left(c_t^h \mid c_{\leq t}^b, c_{<t}^h, a_{\leq t}, I\right).
\end{aligned} \tag{3}
$$

Note that our cross-condition modeling between the body and hand motions makes the most of mutual information in two ways: (1) the current body/hand motions (i.e. c_t^b/c_t^h) depend on past hand/body motion information (i.e. $c_{<t}^b/c_{<t}^h$); (2) we argue that the current body motion c_t^b is also responsible for predicting the distribution of current hand motions. Such modeling guarantees the coherence of the body and hand motions as a whole and thus achieves realistic gestures. Gated PixelCNN [49] is adopted to model these quantities.

Method	Face	
	L2 ↓	LVD ↓
Habibie et al. [23]	0.139	0.257
TalkSHOW (Ours)	0.130	0.248

Method	Body&Hands	
	RS ↑	Variation ↑
Habibie et al. [23]	0.146	0
Audio Encoder-Decoder	0.214	0
Audio VAE	0.182	0.044
Audio+Motion VAE	0.240	0.176
TalkSHOW (Ours)	0.414	0.821

Table 2. Comparison to Habibie et al. [23] and several baselines. ↑ indicates higher is better and ↓ indicates lower is better.

During the training phase, the quantized body/hand motions representation concatenated with the audio and identity features is used for training. A teacher-forcing scheme and cross-entropy loss are adopted for the optimization. At inference, the model predicts multinomial distributions of the future body and hand motions, from which we can sample to acquire codebook indices for each motion. A codebook lookup is then conducted to retrieve the corresponding quantized element of motion, which we feed into the decoder for the final synthesis. Figure 3 (B) illustrates the pipeline. More training details are given in the supplemental material.

5. Experiments

We evaluate the ability of our method in generating body motions (i.e. a sequence of poses) from the speech on the created dataset both quantitatively and qualitatively. Specifically, we choose video sequences longer than 3s and split them into 80%/10%/10% for the train/val/test set. Several metrics are used to measure the realism and diversity of the generated motions including facial expression and hand poses. Furthermore, we conduct perceptual studies to assess the performance of our method.

5.1. Experimental Setup

Evaluation Metrics. Because we model face motion as a deterministic task and the body and hand motions as a non-deterministic task, we assess the generated motion in terms of the realism and the synchronization of face motion, and the realism and the diversity of body and hand motions. Specifically, the following metrics are adopted:

- *L2*: L2 distance between p-GT and generated facial landmarks, including jaw joints and lip shape [39, 64].
- *LVD*: Landmark Velocity Difference calculates the velocity difference between p-GT and generated facial landmarks, which measures the synchronization between speech and facial expression [64].

- *RS*: Score on the realism of the generated body and hand motion. Following [4, 55], we trained a binary classifier to discriminate real samples from fake ones and the prediction represents the realistic score.

- *Variation*: As used in [39], diversity is measured by the variance across the time series sequence of body and hand motions.

Compared Methods. We compare TalkSHOW to Habibie et al. [23], a SOTA speech-to-motion method. Also, we compare several baselines for modeling body and hand motions when using the same face generator as ours:

- **Audio Encoder-Decoder.** It encodes input audio and outputs motions; this is used by [21, 23].

- **Audio VAE.** Given the input audio, the VAE-like structure encodes audio into a Gaussian distribution, and then the sampled audio is fed into the decoder, which transforms the sample into motions.

- **Audio+Motion VAE.** Given the input motion and audio, it adopts a VAE-like structure with two encoders to encode motion and audio into Gaussian distributions, respectively, and then the sampled motion and audio are concatenated and fedinto the decoder for the synthesis.

5.2. Quantitative Analysis

Table 2 shows the comparison results. We see that our method outperforms Habibie et al. [23] across all metrics. Particularly, our method surpasses it in terms of L2 and LVD, which demonstrates the effectiveness of our face generator for generating realistic facial expressions. Also, our method significantly outperforms it in terms of variation, which demonstrates the powerful capacity to generate diverse body and hand motions resulting from our proposed compositional quantized motion representation. Moreover, regarding the realism (RS) for body and hand motion, we surpass Habibie et al. [23] considerably, which confirms the effectiveness of our cross-conditioned autoregressive model in generating realistic motion.

On the other hand, compared to VAE-based models, our method achieves significant gains in both realism and diversity. In particular, we obtain much higher diversity. This indicates the advantage of the learned compositional quantized motion codebooks, which effectively memorize multiple motion modes of the body and hands and thus boost the diversity of the generated body and hand gestures.

5.3. Qualitative Analysis

Figure 4 shows examples of our generated 3D holistic body motion from speech. We see that given the word "*But*" from the speech represents a strengthening tone of voice, our method generates plausible holistic body motions with hands up before saying "*But*" and hands down after saying "*But*". Notably, the generated motions are diverse in many aspects, e.g. the range of motion and which hands to use. For a better illustration of the differences between generated motion samples, we overlap the motions in the same frame with different levels of transparency.

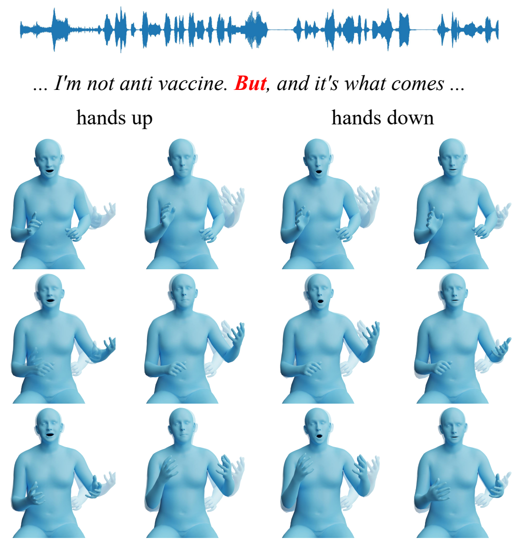

Figure 4. Our method generates diverse motions consistent with the rhythm of the input audio. For instance, we can generate different movements of hands corresponding to the strengthening tone of "*But*" in the speech, e.g. using left hand only (top), right hand only (middle), or both hands (bottom).

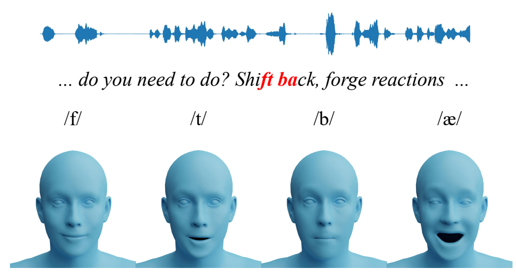

Figure 5. Given speech audio as input, our method generates facial expressions with accurate lip shapes.

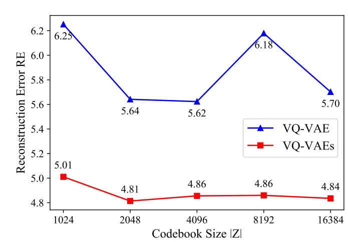

Figure 6. The comparison of VQ-VAE and VQ-VAEs with compositional codebooks.

Figure 5 illustrates the qualitative performance of our face generator. Our approach generates realistic face motions including consistent lip motions with the corresponding phonemes such as /f/, /t/, /b/, and /æ/. Furthermore, our method exhibits a remarkable generalization ability to unseen languages and various audio types, e.g. French and songs. Additional interesting examples can be found in the supplemental video.

5.4. Model Ablation

Effect of Wav2vec Feature in Face Generation. We evaluate the effect of the wav2vec feature used in face generation compared to the MFCC feature. We add an extra encoder to increase the dimension of the MFCC feature from 64 to 256 for a fair comparison. The wav2vec-based model outperforms the MFCC-based model in both metrics (0.130 vs. 0.165 in L2 and 0.251 vs. 0.277 in LVD) due to its larger capacity for modeling the relationship between audio and phonemes. Moreover, we experimentally find that the wav2vec-based model can generalize well to unseen identities; see supplementary for more details.

Effect of Compositional Quantized Motion Codebooks. We analyze the capability of the proposed compositional quantized motion codebooks of VQ-VAEs in efficiently capturing the diverse motion modes represented in motion data, which leads to accurate reconstruction. To this end, we compare VQ-VAE with a single codebook. Reconstruction Error RE is adopted as the metric, in which a lower reconstruction loss indicates a higher capacity. Figure 6 illustrates the results. We see that compared to VQ-VAE with a single codebook, VQ-VAEs with compositional codebooks yield consistently lower RE across different codebook sizes. This demonstrates the effectiveness of the proposed compositional codebooks in modeling the diverse motion modes.

Effect of Cross-Conditional Modeling. In contrast to cross-conditional modeling (w/ c-c), the model without cross condition (w/o c-c) generates body and hand motions independently. Our method w/ c-c yields a higher realistic score than that w/o c-c (0.414 vs. 0.409), benefiting from the cross-conditional modeling between the body and hand motions which leads to more coherent and realistic motions. Note that due to the higher realism, our method w/ c-c attains a slight reduction in diversity (0.821 vs. 0.922 in variance), which is reasonable.

5.5. Perceptual Study

We conduct perceptual studies with Google Forms to evaluate the quality of our reconstruction and generation methods, respectively. We randomly sample 40 videos in total with 10 videos from each speaker. Ten participants took part in the study.

Method	face	body	hands	holistic body
PyMAF-X [62]	0.323	0.500	0.438	0.193
SHOW (ours)	0.898	0.738	0.800	0.768

Table 3. Perceptual study results on reconstruction. For each method, we report the average percentage of answers that the reconstructed results match the input video.

Method	face	body and hands	holistic body
[23] vs. p-GT	0.153	0.141	0.169
TalkSHOW (Ours) vs. p-GT	0.478	0.464	0.458
TalkSHOW (Ours) vs. [23]	0.888	0.910	0.913

Table 4. Perceptual study of motion generation. We use A/B testing and report the percentage of answers where A is preferred over B.

Reconstruction. We assess the quality of our holistic body reconstruct results against PyMAF-X [62], compared with the ground truth. Participants are asked to answer the following questions with Yes or No: Does the reconstructed face/hands/body/full-body match the input video? Table 3 reports the average percentage of answers that the reconstructed results match the input video. We see that our method outperforms PyMAF-X by a large margin.

Motion Generation. We use A/B testing to evaluate our generation results, compared to the p-GT and Habibie et al. [23]. Specifically, participants are asked to answer the following questions with A or B: For the face/body&hands/overall region, which one is a better match with the given speech? Table 4 reports the average preference percentage of answers. We see that participants favor our method over Habibie et al. in terms of all the regions. Not surprisingly, participants perceive the p-GT better over both methods, with our method preferred by many more users.

6. Conclusion

In this work, we propose TalkSHOW, the first approach to generate 3D holistic body meshes from speech. We devise a simple and effective encoder-decoder for realistic face generation with accurate lip shape. For body and hands, we enable diverse generation and coherent prediction with compositional VQ-VAE and cross-conditional modeling, respectively. Moreover, we contribute a new set of accurate 3D holistic body meshes with synchronous audios from in-the-wild videos. The annotations are obtained by an empirical approach designed for videos. Experimental results demonstrate that our proposed approach achieves state-of-the-art performance both qualitatively and quantitatively.

Acknowledgments. We thank W. Zielonka and J. Thies for helping us incorporate MICA into SHOW, C. Ding, H. Jiang, Y. Feng, Z. Liu, and W. Liu for insightful discussions, and B. Pellkofer for IT support. This work was supported, in part, by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B.
Disclosure. https://files.is.tue.mpg.de/black/CoI_CVPR_2023.txt

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 *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1884–1895, 2020. 3

[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 *European Conference on Computer Vision (ECCV)*, pages 248–265. Springer, 2020. 3

[3] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. Style-controllable speech-driven gesture synthesis using normalising flows. In *Computer Graphics Forum (CGF)*, volume 39, pages 487–496. Wiley Online Library, 2020. 3

[4] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5223–5232, 2020. 7

[5] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *arXiv*, 2022. 1, 3

[6] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, pages 12449–12460, 2020. 2, 4

[7] Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In *International Conference on Multimedia (MM)*, pages 2027–2036, 2021. 3

[8] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPl: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, pages 561–578. Springer, 2016. 2

[9] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, 2017. 2, 4

[10] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. Beat: the behavior expression animation toolkit. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 477–486, 2001. 3

[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, 2021. 6

[12] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, pages 20–40, 2020. 2

[13] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael Black. Capture, learning, and synthesis of 3D speaking styles. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10101–10111, 2019. 2, 3

[14] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 4

[15] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. FaceFormer: Speech-driven 3D facial animation with transformers. In *Computer Vision and Pattern Recognition (CVPR)*, pages 18770–18780, 2022. 1, 3, 4

[16] Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool. A 3-D audio-visual corpus of affective communication. *Transactions on Multimedia (TMM)*, 12(6), October 2010. ISSN 1941-0077. 2, 3

[17] Qi Fang, Kang Chen, Yinghui Fan, Qing Shuai, Jiefeng Li, and Weidong Zhang. Learning analytical posterior probability for human mesh recovery. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 2

[18] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, pages 792–804. IEEE, 2021. 2, 4

[19] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3D face

model from in-the-wild images. *Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 4

[20] Ylva Ferstl and Rachel McDonnell. Investigating the use of recurrent motion modelling for speech gesture generation. In *International Conference on Intelligent Virtual Agents (IVA)*, pages 93–98, 2018. 1, 2, 3

[21] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik. Learning individual styles of conversational gesture. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019. 1, 2, 3, 7

[22] Susan Goldin-Meadow. The role of gesture in communication and thinking. *Trends in Cognitive Sciences*, 3 (11):419–429, 1999. 1

[23] 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 *International Conference on Intelligent Virtual Agents (IVA)*, pages 101–108, 2021. 1, 2, 3, 6, 7, 8

[24] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8320–8329, 2018. 2

[25] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018. 2

[26] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 1

[27] Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. Real-time facial surface geometry from monocular video on mobile gpus. *arXiv*, 2019. 4

[28] Adam Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004. 1

[29] Stefan Kopp and Ipke Wachsmuth. Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds*, 15(1):39–52, 2004. 3

[30] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. Analyzing input and output representations for speech-driven gesture generation. In *International Conference on Intelligent Virtual Agents (IVA)*, pages 97–104, 2019. 3

[31] 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 *International Conference on Multimodal Interaction (ICMI)*, pages 242–250, 2020. 3

[32] Sergey Levine, Christian Theobalt, and Vladlen Koltun. Real-time prosody-driven synthesis of body language. In *ACM SIGGRAPH Asia*, pages 1–10, 2009. 1

[33] Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. Gesture controllers. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 1–11, 2010. 1, 3

[34] Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. NIKI: Neural inverse kinematics with invertible neural networks for 3D human pose and shape estimation. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 2

[35] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 11293–11302, 2021. 3

[36] Miao Liao, Sibo Zhang, Peng Wang, Hao Zhu, Xinxin Zuo, and Ruigang Yang. Speech2Video synthesis with 3D skeleton regularization and expressive body poses. In *Asian Conference on Computer Vision (ACCV)*, 2020. 3

[37] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 10462–10472, 2022. 3

[38] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. Virtual character performance from speech. In *Symposium on Computer Animation (SCA)*, pages 25–35, 2013. 1

[39] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Computer Vision and Pattern Recognition (CVPR)*, pages 20395–20405, 2022. 6, 7

[40] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas,

and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2, 3, 4

[41] Isabella Poggi, Catherine Pelachaud, F de Rosis, Valeria Carofiglio, and B De Carolis. Greta. A believable embodied conversational agent. In *Multimodal Intelligent Information Presentation*, pages 3–25. Springer, 2005. 3

[42] Shenhan Qian, Zhi Tu, Yihao Zhi, Wen Liu, and Shenghua Gao. Speech drives templates: Co-speech gesture synthesis with learned templates. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11077–11086, 2021. 3

[43] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. MeshTalk: 3D face animation from speech using cross-modality disentanglement. In *International Conference on Computer Vision (ICCV)*, pages 1173–1182, 2021. 1, 3

[44] Najmeh Sadoughi, Yang Liu, and Carlos Busso. Meaningful head movements driven by emotional synthetic speech. *Speech Communication*, 95:87–99, 2017. 1

[45] Md Sahidullah and Goutam Saha. Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. *Speech Communication*, 54(4):543–565, 2012. 5

[46] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 11050–11059, 2022. 3

[47] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J Black. TRACE: 5d temporal regression of avatars with dynamic cameras in 3d environments. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 2

[48] Kenta Takeuchi, Souichirou Kubota, Keisuke Suzuki, Dai Hasegawa, and Hiroshi Sakuta. Creating a gesture-speech dataset for speech-based automatic gesture generation. In *International Conference on Human-Computer Interaction (HCI)*, pages 198–202. Springer, 2017. 2, 3

[49] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelCNN decoders. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 29, 2016. 6

[50] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. 2, 5

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. 4

[52] Petra Wagner, Zofia Malisz, and Stefan Kopp. Gesture and speech in interaction: An overview. *Speech Communication*, 57:209–232, 2014. 1

[53] Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Alexander Hypes, Taylor Koska, Steven Krenn, Stephen Lombardi, Xiaomin Luo, Kevyn McPhail, Laura Millerschoen, Michal Perdoch, Mark Pitts, Alexander Richard, Jason Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shoou-I Yu, and Yaser Sheikh. Multiface: A dataset for neural face rendering. *arXiv*, 2022. 2, 3

[54] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6184–6193, 2020. 2

[55] Jing Xu, Wei Zhang, Yalong Bai, Qibin Sun, and Tao Mei. Freeform body motion generation from speech. *arXiv*, 2022. 1, 2, 3, 7

[56] Payam Jome Yazdian, Mo Chen, and Angelica Lim. Gesture2Vec: Clustering gestures using representation learning methods for co-speech gesture generation. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 3100–3107. IEEE, 2022. 3

[57] Hongwei Yi, Chen Li, Qiong Cao, Xiaoyong Shen, Sheng Li, Guoping Wang, and Yu-Wing Tai. MMFace: A multi-metric regression network for unconstrained face reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7663–7672, 2019. 2

[58] Hongwei Yi, Chun-Hao P. Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J. Black. Human-aware object placement for visual environment reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 4

[59] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *International Conference on Robotics and Automation (ICRA)*, pages 4303–4309. IEEE, 2019. 2, 3

[60] 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. *Transactions on Graphics (TOG)*, 39(6):1–16, 2020. 2, 3

[61] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In *International Conference on Computer Vision (ICCV)*, pages 11446–11456, 2021. 2

[62] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. PyMAF-X: Towards well-aligned full-body model regression from monocular images. *arXiv*, 2022. 2, 4, 8

[63] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. VisemeNet: Audio-driven animator-centric speech animation. *Transactions on Graphics (TOG)*, 37(4):1–10, 2018. 3

[64] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. MakeItTalk: Speaker-aware talking-head animation. *Transactions on Graphics (TOG)*, 39(6):1–15, 2020. 6

[65] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision (ECCV)*, 2022. 4