

SpeechAct: Towards Generating Whole-body Motion from Speech

Jinsong Zhang[†], Minjie Zhu[†], Yuxiang Zhang, Zerong Zheng, Yebin Liu, *Member, IEEE*,
Kun Li*, *Senior Member, IEEE*

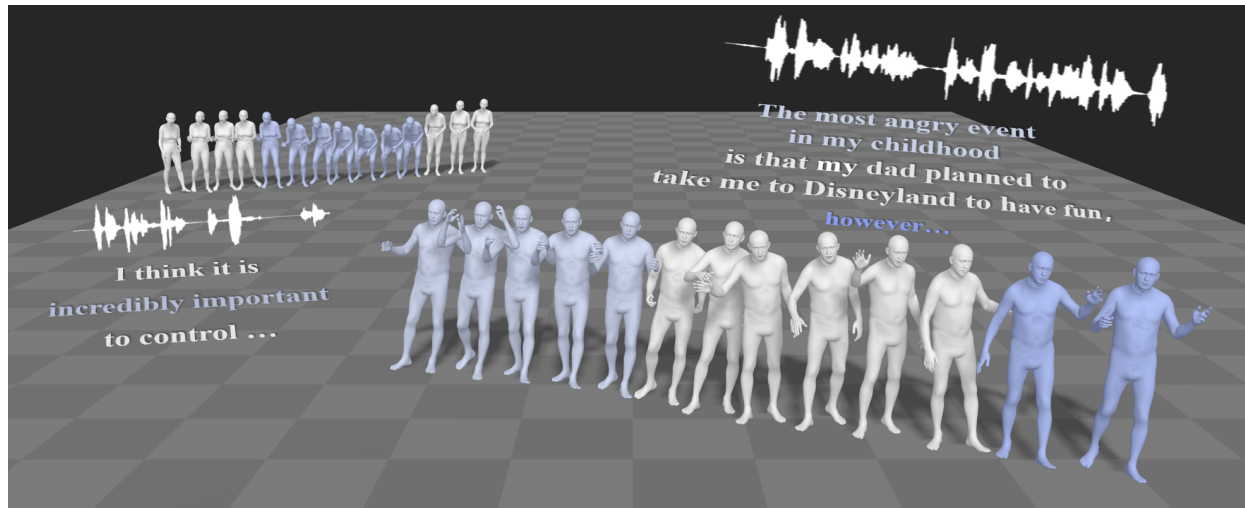


Fig. 1: Given an audio input, our model can generate natural and diverse human motion sequences. There are two samples that are uniformly sampled from generated space. The human body meshes corresponding to the same text color indicate the motion generated by the driven speech content.

Abstract—Whole-body motion generation from speech audio is crucial for computer graphics and immersive VR/AR. Prior methods struggle to produce natural and diverse whole-body motions from speech. In this paper, we introduce a novel method, named SpeechAct, based on a hybrid point representation and contrastive motion learning to boost realism and diversity in motion generation. Our hybrid point representation leverages the advantages of keypoint representation and surface points of 3D body model, which is easy to learn and helps to achieve smooth and natural motion generation from speech audio. We design a VQ-VAE to learn a motion codebook using our hybrid presentation, and then regress the motion from the input audio using a translation model. To boost diversity in motion generation, we propose a contrastive motion learning method according to the intuitive idea that the generated motion should be different from the motions of other audios and other speakers. We collect negative samples from other audio inputs and other speakers using our translation model. With these negative samples, we pull the current motion away from them using a contrastive loss to produce more distinctive representations. In addition, we compose a face generator to generate deterministic face motion due to the strong connection between the face movements and the speech audio. Experimental results validate the superior performance of our model. The code will be available for research purposes.

Index Terms—Speech-driven motion generation, hybrid point representation, contrastive motion learning, VQ-VAE.

1 INTRODUCTION

HUMAN motion generation from speech audio [1]–[3] is a critical area in computer graphics and immersive VR/AR [4]–[7], which has been extensively studied as a way of human behavior understanding [1]. Given a speech

recording, the goal is to generate a spectrum of diverse yet natural motion sequences, which is in line with real-life scenarios and can meet the varying needs of different individuals.

Existing approaches concentrate on translating speech to the motion of the partial body [3], [8]–[10]. They use keypoints of the body as their motion representation, which is easy to learn and contains local details for hands. However, the keypoints leave degrees of freedom undefined, *e.g.*, the rotation of a limb, leading to inaccurate and unrealistic results when fitting or animating a full 3D body. For the whole-body motion generation, TalkShow [1] adopts the parametric representation of SMPL-X body mesh as the

- [†] Equal contribution.
- * Corresponding author: Kun Li (Email: lik@tju.edu.cn)
- Jinsong Zhang, Minjie Zhu, and Kun Li are with the College of Intelligence and Computing, Tianjin University, Tianjin 300350, China.
- Yuxiang Zhang and Yebin Liu are with the Department of Automation, Tsinghua University, Beijing 100084, China.
- Zerong Zheng is with NNKosmos Technology, Hangzhou, P.R.China.

motion representation, and proposes a two-stage model to recover the full 3D body. However, we observe that the motion generated by TalkShow often contains discontinuous results, for example, jitters and disharmony contact between the feet and the ground. This is because they use parametric representation defined in the axis-angle space using the kinematic tree, which may introduce the discontinuity caused by the axis-angle representation, leading to the difficulty of modeling continuous latent space [11]–[13]. Moreover, due to the inherent one-to-one mapping in our real life and training pairs, all the above approaches tend to learn an average motion lacking diversity. However, in fact, given a speech recording, different speakers tend to exhibit varying motions in different situations. Therefore, how to generate diverse results is also a challenging problem.

In this work, to generate smooth and natural motion, we propose a hybrid point representation for whole-body motion generation from speech audio. Specifically, the keypoint representation is easy to learn and contains local details for hands, but lacks surface constraints to obtain the 3D body model. To combat this problem, we adopt the surface points of SMPL-X body mesh as a part of our hybrid point representation to eliminate the ambiguity due to the keypoint representation. Therefore, defined in Euclidean space, our hybrid point representation is easy to learn and encompasses global constraints and local details for whole-body motion generation. This helps generate smooth and natural results, *e.g.*, avoiding foot skating. Besides, to provide an easy-to-use output representation that can be used in many applications, *i.e.*, SMPL-X body mesh, we design a generator to recover the parameters of SMPL-X from our hybrid representation.

With this hybrid point representation, we can generate smooth and natural results, however, similar to previous approaches, the diversity of generated results is still limited due to the one-to-one mapping in the training data. This one-to-one mapping in the training data makes the model capture similar motions, *e.g.*, certain habitual body motions of a speaker and similar body motions of different speakers due to specific meanings, but this limits the model to generate diverse motions. To boost the diversity of generated results, an intuitive idea is that, the generated motion of a specific audio should be different from the generated motions of other audios and other speakers. To achieve this, we introduce a novel contrastive motion learning method to obtain distinctive motion representations. Specifically, we first collect negative samples that are discrete motion representations from other audio inputs and other speakers using the translation model. Then, we pull the current generated motion away from the negative samples by a contrastive loss, and thus the generated motion can be more distinctive. By using the target motion as a positive sample, our approach encourages diverse motion learning while capturing similar motions.

Built on our hybrid point representation and motion contrastive learning method, we present a novel model, named SpeechAct, to generate natural and diverse body motion. This model first constructs a quantized motion codebook based on our presentation, and then regresses the distinct motion from speech audio using our contrastive motion learning. In addition, to support whole-body motion synthe-

sis, we compose a face generator to generate deterministic face motion due to the strong connection between the face movements and the speech audio. Experimental results compared with several state-of-the-art methods validate the effectiveness of our model. We give an application that animates avatars using speech inputs, which can be applied in immersive AR/VR [14], [15]. Moreover, our model can generate promising results conditioned on different languages. Fig. 1 presents two samples generated by our model. The human body meshes corresponding to the same text color indicate the motion generated by the driven speech content. It can be seen that our approach not only generates smooth sequences of movements containing diverse poses, but also ensures that these sequences correspond to the audio content.

Our main contributions can be summarized as follows:

- We design SpeechAct, a novel framework based on a new representation to generate whole-body motion from speech audio, which can produce more natural and diverse results.
- We introduce a novel hybrid point representation that contains global constraints and local details for whole-body, and a generator that can recover parameters of SMPL-X body mesh, which can be used for avatar animation.
- We propose a contrastive motion learning method to learn a more distinctive motion representation, which improves the ability of our model to generate diverse results.
- Experimental results demonstrate our model can generate natural and diverse results. Our method can also be generalized to other languages.

2 RELATED WORK

2.1 Human Motion Generation

Human motion generation is the task of synthesizing realistic and natural human movements for various applications, such as animation, gaming, robotics, and VR/AR. The data representations can be divided into two kinds, *i.e.*, keypoint-based and rotation-based. The keypoint-based representation [3], [16], [17] adopts a set of keypoints that are specific points on the body, *e.g.*, joints. The keypoints are defined using 2D/3D coordinates in the pixel or world coordinate system. Some works [13], [18] use sparse markers to represent human pose, which can provide more comprehensive information for rotation. However, the keypoint-based methods can provide local details for hands, but can not constrain the global shape, such as the rotation of limbs, which can lead to unrealistic results when fitting or animating a full 3D body. The rotation-based methods [19], [20] adopt the rotation of the body part relative to their parent, which helps to constrain the body pose by modeling the body, face and hands jointly. However, this representation can generate discontinuous results due to its specific space [11]–[13], [21], *e.g.*, axis-angle space for SMPL-X body model.

In this paper, we propose a hybrid point representation for whole-body motion generation. Our hybrid point representation is defined in Euclidean space and is easy to learn, which encompasses global constraints and local details.

2.2 Whole-body Motion Generation from Speech

Previous approaches focus on generating different parts of whole-body from speech, *e.g.*, face, hands and body. Existing speech-driven 3D face animation methods [22], [23] rely on 4D face scan datasets to train their models. For body motion generation, two kinds of methods have been explored: rule-based methods and learning-based methods. Rule-based methods [24]–[26] convert the input speech to motions from a pre-collected motion database according to manually designed rules. These methods are interpretable and controllable, but are difficult to generate complex and realistic motions. Due to limitations in existing datasets [3], [8], [27], [28], previous learning-based methods primarily concentrate on generating partial human body motions from speech [3], [29]–[32]. Most of them adopt keypoints as motion representation, achieving accurate results closely approximating the ground-truth keypoints but leading to inaccurate and unrealistic results when fitting or animating a full 3D body. Besides, these methods generate deterministic results. This means that they fail to generate diverse results given the same audio signal, which is inconsistent with cognitive understanding.

Some approaches cooperate GANs [7], [33]–[35], VAEs [6], [36], [37], VQ-VAEs [9], [38], and normalizing flows [39], [40] to increase the diversity of the generated results, but the results are inadequate [1]. Recently, due to the great success of diffusion models [41], some methods [2], [42]–[44] introduce diffusion models to generate human motion from speech, which can generate promising results. However, this kind of method requires a large number of diffusion steps, and thus has low generation efficiency. Besides, modeling part of the whole-body is insufficient for a comprehensive understanding of human behavior. To generate whole-body motion, most similar to our work, Yi *et al.* [1] fit the SMPL-X body mesh based on the speaker-specific gesture dataset [45], but the motions in their dataset focus on face and upper-body gestures. Based on their dataset, they propose to generate diverse motions for the body and hands using two independent codebooks. Liu *et al.* [46] propose a new whole-body co-speech dataset, *i.e.*, BETAX dataset, and design a masked audio gesture transformer to regress the gesture from the audio input. However, because of the inherent one-to-one mapping between input and output, the model struggles to generate natural and diverse results. Besides, they adopt parameters of SMPL-X body mesh as the motion representation, which introduces complexity for the prediction task [47] and can lead to discontinuous problems.

In this paper, we introduce a novel hybrid representation by combining keypoints and surface points of SMPL-X body mesh, which helps to generate continuous results and is easy to expand to other applications. Besides, we introduce a novel contrastive motion learning method by collecting negative motion samples and pulling the generated motion away from them, which promotes the model to generate more diverse results.

2.3 Contrastive Learning

Contrastive learning is a powerful self-supervised learning method, which has achieved great progress in many computer vision and computer graphic tasks. Traditionally, the

central idea is to take a sample as an anchor, compare it with other samples, and pull it closer to positive samples while pushing it apart from negative samples. By doing so, we can get a meaningful representation. Though designed for 2D image tasks [48], [49], contrastive learning has been used in 3D tasks [50], [51] and multi-modal tasks [42], [52]. Ao *et al.* [42] adopt contrastive learning to learn a gesture-transcript joint embedding space to align the gesture space and the transcript space. Zhao *et al.* [52] use a similar idea to achieve music-driven dance generation.

Different from previous methods, instead of aligning the multi-modal information to obtain a joint embedding space, we propose a contrastive motion learning method to boost the diversity of generated results.

3 METHOD

Given a speech recording, our work aims to generate diverse whole-body motion sequences that are in harmony with the provided speech. We design a novel framework, named SpeechAct, to achieve this. The most significant differences with existing approaches are that, to generate natural and diverse results for the body and hands, we propose a hybrid point representation to form our motion efficiently, and a contrastive motion learning method to distinguish the generated motion from other motions to boost the diversity of the generated results. Fig. 2 shows the overall framework of our SpeechAct. Our model consists of two generators: a body generator (Sec. 3.1) and a face generator (Sec. 3.2). Because the audio signal is not very closely related to hand and body gestures [1], the body generator aims to generate natural and diverse results. For an effective and versatile motion representation, we introduce a hybrid point representation (Sec. 3.1.1) and develop a generator to obtain SMPL-X model parameters. To ensure the generation of coherent and diverse results, we leverage advancements in VQ-VAE to learn a quantized motion space (Sec. 3.1.2). Building upon this quantized motion space, we propose a translation model (Sec. 3.1.3) and introduce a novel contrastive motion learning method, allowing us to transform audio signals into motion representations while promoting diversity in body and hand motions. In contrast, speech content, especially phonetic information, is closely related to face motion, notably lip motion, thus the face generator is designed using an encoder-decoder architecture and is responsible for generating deterministic results.

3.1 Body Generator

For the body generator, we aim to generate natural and diverse motion sequences for the body and hands given an audio input. Previous approaches [3], [8] adopt keypoints for motion representation, limiting their ability to fully reconstruct or animate a 3D body. TalkShow [1] utilizes SMPL-X model parameters, resulting in discontinuous motion sequences. To generate more realistic and continuous results, we propose a hybrid point representation with a generator, which can generate continuous results efficiently and deliver the SMPL-X model accurately. Besides, because the input and the output are one-to-one mappings, previous methods struggle to generate natural yet diverse results

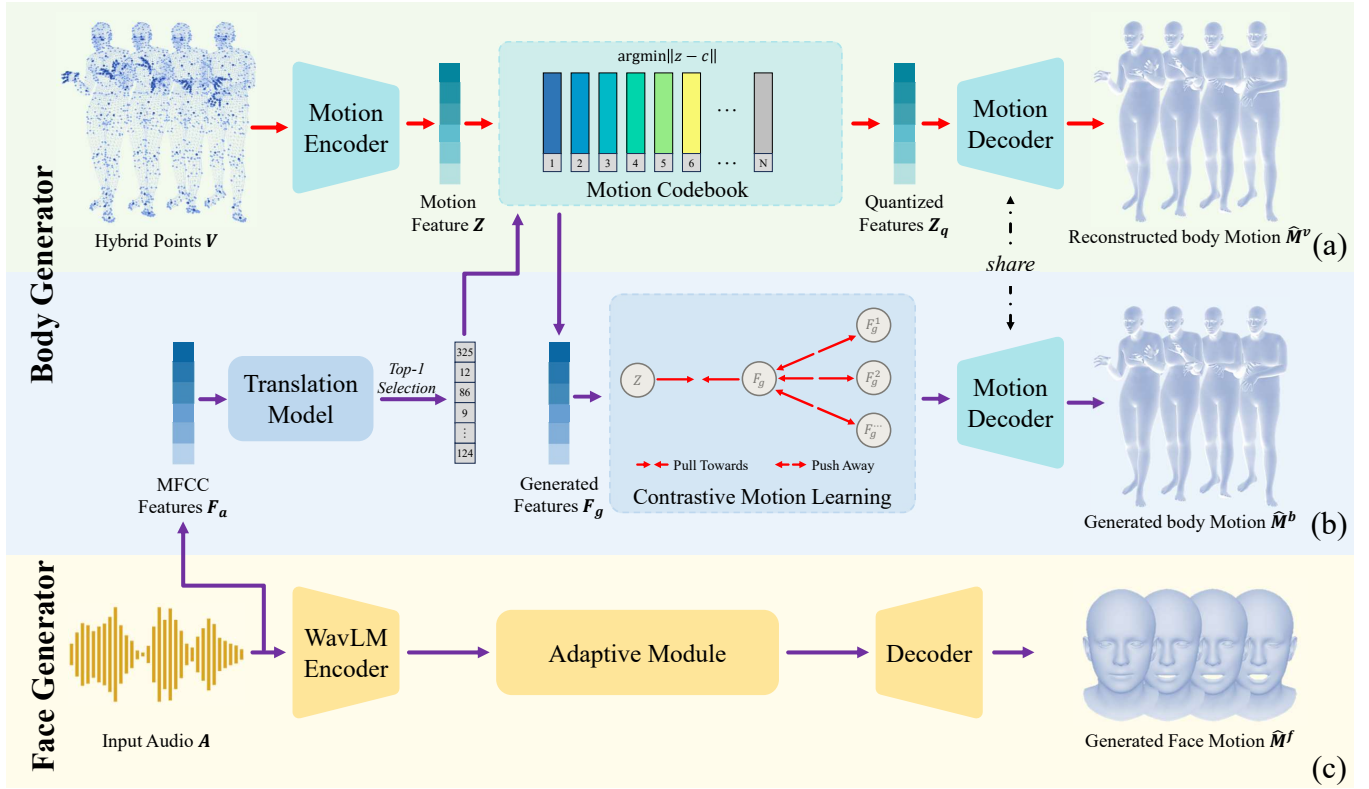


Fig. 2: The detailed architecture of SpeechAct. To generate whole-body motion, our model includes a two-stage body generator to generate diverse motions for the body and hands and a face generator to output deterministic results. Specifically, our model includes: (a) a VQ-VAE based on our proposed hybrid point representation to learn a motion codebook, (b) a translation model with a contrastive motion learning method to generate diverse motion codes from the learned motion codebook, and (c) an encoder-decoder architecture to generate deterministic face motion. The red lines indicate these modules are used for training, and the purple lines mean that these modules are applied for both training and inference.

due to the challenges of balancing constraints between precision and diversity. To alleviate this problem, we design a two-stage model and introduce a novel contrastive motion learning method to boost the diversity of generated results. Fig. 2 shows the framework of our body generator. Specifically, we first learn a motion codebook by encoding motion sequences into quantized codes. To synthesize more natural hand motion, we model the body and hands separately using three different codebooks. Subsequently, our translation model autoregressively transforms audio signals into motion sequences. With the contrastive motion learning method, the translation model can generate more distinctive motion representations, which helps to generate diverse results.

3.1.1 Hybrid Point Representation

As shown in Fig. 3, our hybrid points representation consists of two vital components: surface points v_s (the points in light blue) and keypoints v_k (the points in dark blue). Surface points are extracted from the SMPL-X body mesh using mesh sampling operation [53], serving as a global constraint to the freedom of the body. On the other hand, keypoints, which are defined by SMPL-X body mesh and can be obtained using the joint regressor from SMPL-X [54], capture fine-grained details, particularly in areas like the hands, enriching the representation. In practice, we adopt

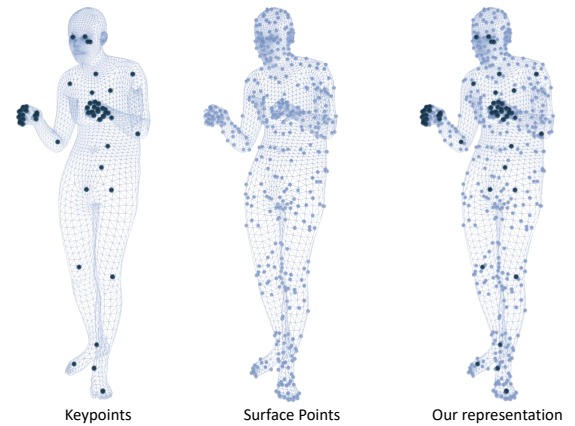


Fig. 3: The overview of our hybrid representation

all keypoints, *i.e.*, 55 keypoints, of SMPL-X body mesh, and use 431 points sampled from the surface. As both keypoints and surface points are defined in Euclidean space, the final representation $v \in \mathcal{R}^{486 \times 3}$, results from the concatenation of v_s and v_k , with each element representing a 3D location. This combined representation facilitates the capture of both global constraints and local details. This representation is employed to represent a motion sequence, denoted as

$V = \{v_t\}_{t=0}^T$, where T represents the number of frames in a motion sequence. It encapsulates the dynamic evolution of the hybrid point-based representation over time, enabling comprehensive modeling and analysis of human body motion.

Inspired by [11], to eliminate the gap between the SMPL-X body mesh and our representation, we design a simple generator \mathcal{G}_P , where the input is v_t and the output is the parameters of the SMPL-X body mesh. This generator \mathcal{G}_P consists of a series of neural network layers, specifically a stack of three residual blocks, each with varying input and output channels. These layers are responsible for transforming the input point representation into a more meaningful intermediate feature representation. Subsequently, a final convolutional layer is employed to produce the desired output, which represents the parameters of the SMPL-X body mesh. By doing so, even training the body and face separately, our model can ensure consistency between body and face to generate seamless and continuous results. This generator is trained with our body generator, which is represented in Sec. 3.1.2.

Discussion. Our hybrid point representation has two advantages compared with previous approaches:

- Compared to the keypoint representation, our hybrid representation can eliminate the ambiguities when recovering the shape and pose of the whole-body. Besides, with the generator \mathcal{G}_P , our representation can be transformed to SMPL-X body mesh easily, and this process is differentiable.
- Our representation is defined in the Euclidean space. This choice not only facilitates the generation of continuous motion, *e.g.*, avoiding foot skating, but also enhances the model's learning process compared to parameter-based representations.

3.1.2 Quantized Motion Modeling

As shown in Fig. 2 (a), the first stage of our body generator is quantized motion modeling. Given a set of motion sequences $\mathcal{M} = \{M_k\}_{k=0}^K$, where K is the number of motion sequences, we first extract our hybrid point representation $\mathcal{V} = \{V_k\}_{k=0}^K$. With a sample V from \mathcal{V} , we adopt a Vector Quantized-VAE (VQ-VAE) [55] to learn a meaningful and compact motion space. We first obtain motion features $Z \in \mathcal{R}^{\frac{T}{w_m} \times n_z}$ from the input motion sequence, where T is the number of frames in V , n_z is the channel dimension of quantized features and w_m is the temporal window size.

Denote the learnable motion codebook as $C = \{c_i\}_{i=1}^N$, where N is the length of the codebook. Then, we quantize the motion features by replacing them with the nearest codes in the codebook. Specifically, for each row z_t in Z , the quantized feature can be obtained by:

$$c_t = \arg \min_{c_i \in C} \|z_t - c_i\|_2. \quad (1)$$

Therefore, the quantized features $Z_q = \{c_t\}_{t=0}^{\frac{T}{w_m}}$. Afterward, the reconstructed motion \hat{V} is delivered by a decoder. To output the SMPL-X body mesh, we cooperate with the generator \mathcal{G}_P in the motion decoder, which can transform our hybrid point representation into SMPL-X body mesh.

We train this VQ-VAE with the reconstruction loss and the vector quantization loss to obtain a meaningful motion space. The reconstruction loss is used to recover the motion sequence accurately. Following [56], we recover the locations, velocities, and accelerations of the movements, mathematically:

$$\mathcal{L}_{rec} = \|\hat{V} - V\|_1 + \alpha_1 \|\hat{V}' - V'\|_1 + \alpha_2 \|\hat{V}'' - V''\|_1, \quad (2)$$

where V' and V'' are the first-order and second-order partial derivatives of motion representations V , and α_1 and α_2 are the balancing weights of the corresponding items.

The vector quantization loss can be written as:

$$\mathcal{L}_{vq} = \|\text{sg}[Z] - Z_q\|_2 + \beta \|Z - \text{sg}[Z_q]\|_2, \quad (3)$$

where $\text{sg}[\cdot]$ is the stop gradient function [57], the first item is the codebook loss and the second item is the commitment loss with the weight β . The codebook loss and the commitment loss are used to align the vector space of the codebook and the outputs of the encoder. The codebook loss lets the quantized features from the codebook close to the outputs of the encoder, and the commitment loss brings the outputs of the encoder close to the codes in the codebook.

In practice, we adopt three codebooks, *i.e.*, left-hand codebook, right-hand codebook, and body codebook, to model more detailed motions for the body and hands, respectively. Therefore, the encoded feature Z can be split into three parts Z^{lh}, Z^{rh}, Z^b according to different body parts of our hybrid representation before quantization, and these three features can be quantized using Eq. 1. The quantized feature Z_q is the concatenation of left-hand quantized feature Z_q^{lh} , right-hand quantized feature Z_q^{rh} , and body quantized feature Z_q^b . Therefore, though we adopt three codebooks, they can be optimized using Eq. 2 and Eq. 3 simultaneously.

To train the generator \mathcal{G}_P simultaneously, we adopt the L1 distance to regress the parameters of SMPL-X body mesh, mathematically:

$$\mathcal{L}_{sx} = \|\mathcal{G}_P(\hat{V}) - \theta\|_1, \quad (4)$$

where θ is the ground-truth parameters of SMPL-X body mesh.

3.1.3 Translation Model

With the learned motion codebook, as shown in Fig. 2 (b), we aim to translate the audio signal into the motion representation from the motion codebook, and then output the motion sequence using the learned motion decoder in the second stage. Speech content, especially phonetic information, is closely related to face motion, notably lip motion, while its link to body/hand motion is weaker. Compared with phonetic information in the audio signal, speech rhythm and beat are more related to body and hand motion. Therefore, given the input audio signal $A = \{a_t\}_{t=0}^T$, we first transform it to 64-dimensional Mel-Frequency Cepstral Coefficients (MFCC) features F_a containing rhythm and beat information. To regress the motion representation with speaker information, we design a translation model \mathcal{T} to translate the input audio signal into the motion codes. Specifically, we concatenate the speaker embedding, *i.e.*, a one-shot embedding, with F_a , and employ several retention

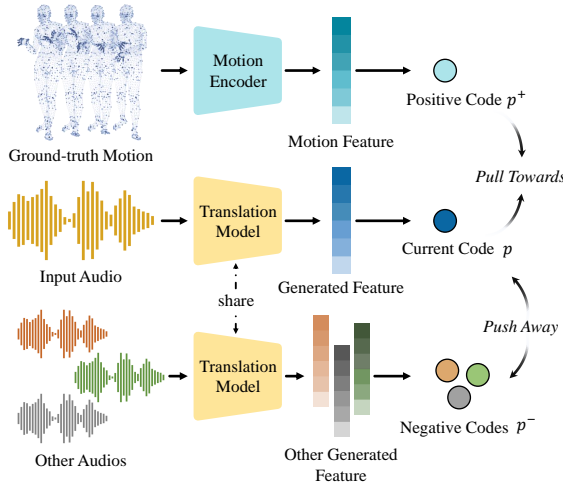


Fig. 4: Details of contrastive motion learning. We take the quantized features from the ground-truth motion as the positive sample, and the generated features from other audios as the negative samples. By pulling away the current generated feature from the negative samples, we can obtain more distinctive representations.

blocks [58] to extract the regressed features $F_r \in \mathcal{R}^{\frac{T}{w_m} \times N}$, where w_m is the temporal window size and N is the length of the motion codebook. Subsequently, we adopt a SoftMax layer to obtain the regressed logits, and select the highest score at each temporal window as the index of the quantized feature. Thus, we can obtain the generated features F_g by selecting features according to the selected indexes from the learned motion codebook. With the learned motion decoder, we can obtain natural results in harmony with the input audio. To regress natural results, we adopt cross-entropy loss \mathcal{L}_{ce} to optimize the translation model. As an auto-regressive model, the translation model can generate motions of arbitrary length by adjusting the input sequence of previously predicted tokens. Besides, during inference, we randomly initialize the first token using the index of the codebook to generate diverse results.

Contrastive Motion Learning. Because the input audio and the output motion are one-to-one mappings in training pairs, the diversity of the generated results is still limited. An intuitive idea is that, given a speech recording, the motion for a specific speaker can be different from other speakers, and can be different from the motions driven by other speech recordings. To achieve this, we introduce a novel contrastive motion learning method, which can distinguish the current generated motion from other motions driven by different audios or generated by other speakers. As shown in Fig. 4, different from previous works [9], [42] using contrastive loss to align multi-modal information, *e.g.*, text and motion, we take the early generated quantized features from other speakers and other speech recordings as the negative samples, and the ground-truth quantized features as the positive sample to boost the diversity of generated results. Specifically, we adopt global average pooling along time dimension to obtain motion codes as the negative or positive samples. Evidently, the positive sample is high-quality and in harmony with the input audio signal, while

the negative samples are low-quality and have no connection with the input audio signal. Let p denote the current generated features F_g , p^+ that of ground-truth quantized feature from V , and $\{p_t^-\}_{t=1}^L$ the generated features from L negative samples. To boost the diversity of the generated results, our goal is to ensure that the generated features are different from the negative samples, *i.e.*, minimizing mutual information. Therefore, the contrastive loss can be formulated as:

$$\mathcal{L}_{cm} = -\log \frac{\exp(\frac{p^T p^+}{\tau})}{\exp(\frac{p^T p^+}{\tau}) + \sum_{j=1}^L \exp(\frac{p^T p_j^-}{\tau})}, \quad (5)$$

where τ is set to 0.7 and L is set to 1024 in our experiments. By doing so, our model can produce more realistic and diverse motions while maintaining fidelity to ground-truth motions.

3.2 Face Generator

In this work, our goal is to generate whole-body motion from speech audio, and the most important part is body generation due to its diverse character. To support the overall whole-body motion synthesis, we include a face generator to generate facial motion from speech audio. Because the face is closely related to the input signal, especially phonetic information, as shown in Fig. 2 (c), we adopt an encoder-decoder architecture to regress the facial movement. Specifically, we first encode the audio signal $A = \{a_i\}_{i=0}^T$ using a pre-trained WavLM model [59], where the feature extraction is adapted to generate a representation suitable for downstream tasks. This representation is concatenated with speaker embedding to obtain speaker information, and then refined by an adaptive module to extract and aggregate the audio features adaptively. The adaptive module consists of a residual block, followed by an attention block, and then two more residual blocks. This hierarchical arrangement of convolutional and attention mechanisms aims to robustly capture and understand the speaker-specific characteristics and the speech content to generate more reliable results. Subsequently, a series of residual blocks with attention mechanisms are used to decode the audio representation, transforming it into a deterministic face motion sequence. The detailed architecture can be found in the supplemental document. It should be noted that the face generator serves as a complementary aspect and can be replaced by state-of-the-art facial motion synthesis models [60]–[63]. The face generator is trained with face reconstruction loss: L1 distance for expressions and L1 distance for the first order of the face motion representations.

3.3 Implementation Details

We conduct our experiments on a desktop with a GeForce RTX 3090 GPU. For the face generator, we adopt the parametric representation for deterministic results, and the parameters of WavLM are frozen. For the body generator, we adopt our proposed hybrid points representation. We first train the VQ-VAE to learn the motion codebook, and the generator from our representations to obtain parameters of SMPL-X. The weights $\alpha_1, \alpha_2, \beta$ of different losses are set to 0.5, 0.5, 0.25. We set the channel dimensions of quantized

features Z_q^{lh} , Z_q^{rh} , Z_q^b as 128, 128, 512. The length of the temporal window size is set to $\frac{1}{30}$ s for each input audio. Then, we train the translation model and freeze the parameters of the motion decoder. For all training models, we adopt the Adam [64] optimizer with coefficients 0.9 and 0.999 for computing running averages of gradients and their squares. The learning rates for different stages are all set to $1e-4$.

4 EXPERIMENTAL RESULTS

4.1 Experimental Setting

Dataset. To train and evaluate our model, we adopt BEAT2 dataset [46] as the benchmark to conduct experiments. The BEAT2 dataset contains 60 hours captured from 25 speakers in English, which is split into BEAT2-Standard (27 hours) and BEAT2-Additional (33 hours) based on the type of speech and conversation sections. We split BEAT2-Standard into 80%, 10%, and 10% split for the train/val/test set.

Baselines. To validate the performance of our model, We first compare with TalkShow [1], which is the most related work that generates whole-body motion from speech. Besides, we also compare with three state-of-the-art methods: Audio2Gesture [36], DiffStyle [2] and EMAGE [46]. Because Audio2Gesture and TalkShow are trained using other part body dataset, we re-train them on BEAT2 dataset using the public codes. Besides, we implement two baselines for generating body and hand motions to validate the effectiveness of our model. All the experiments are conducted on the same training and testing set. It should be noted that DiffStyle and EMAGE are tested on 1 speaker. Therefore, we adopt the same speaker to evaluate their performance additionally for a fair comparison with them.

- **Audio Enc-Dec.** Given a speech recording, this model first encodes the input and then decodes the features to deliver the motion, which is the same as [45].
- **Motion VAE.** In this model, we first train a VAE using motion sequences to learn a latent space. Then, given a speech audio, we encode the audio signal and concatenate it with a sampling latent code from latent space to output the generated motion using a decoder.

Metric. For the body and hands, we aim to generate natural and diverse results in harmony with the input audio. Therefore, we adopt several evaluation metrics to validate the diversity and realism of generated results.

- **FID:** We calculate Fréchet Inception Distances (FID) [65] using kinetic features [66] (FID-k) and geometric features (FID-g) to measure the quality of generated results. Kinetic features are defined by motion velocities and energies, while geometric features are defined by joint angles and relative joint positions.
- **BeatAlign:** As used in [56], we employ chamfer distance between beats of the input audio and movements of the body and hands to measure the harmony between them, which indicates the quality of generated results.
- **Div-in:** Following [67], we evaluate the diversity of motions corresponding to an individual speech

TABLE 1: Quantitative comparison of body motion on BEAT2 dataset with several baselines. Ours* means we evaluate our method on the single speaker for a fair comparison with DiffStyle and EMAGE. Note that we report $FSR \times 10^{-2}$.

Model	FID-k ↓	FID-g ↓	BeatAlign ↑	Div-in ↑	Div-out ↑	FSR ↓
Audio Enc-Dec	10.36	8.907	0.5504	0.8632	0	2.99
Motion VAE	10.73	8.722	0.5363	0.8733	0.0571	6.78
Audio2Gesture	20.43	4.826	<u>0.5572</u>	0.3893	1.312	1.26
TalkShow	14.37	<u>4.535</u>	0.5479	<u>0.8966</u>	<u>1.744</u>	<u>0.61</u>
Ours	3.839	4.359	0.5603	0.9700	2.264	0.38
DiffStyle	44.91	<u>6.576</u>	0.5811	1.539	<u>0.3340</u>	<u>1.69</u>
EMAGE	<u>6.620</u>	10.12	0.5487	<u>2.817</u>	0	2.28
Ours*	5.979	5.505	<u>0.5704</u>	3.335	7.840	1.19

TABLE 2: Quantitative comparison of face motion on BEAT2 dataset. We report $LVD \times 10^{-2}$ and $LD \times 10^{-2}$.

Model	LVD ↓	LD ↓
TalkShow [1]	4.158	0.8731
EMAGE [46]	4.359	0.9546
Ours	3.829	0.7896

audio through variations in body poses across the temporal sequence, which is denoted as Div-in.

- **Div-out:** Given a specific speech audio, diverse results for the same speaker are expected. To evaluate this, for all audio clips in our test set, we calculate the average L2 distance between the two motion clips generated from an audio clip, denoted as Div-out.
- **FSR:** With our hybrid representation, our model can generate smooth results alleviating foot skating. To validate this character, following [44], [68], we identify frames where the foot slides beyond a certain distance while in contact with the ground (i.e., foot height < 5 cm) as foot-skating frames. We report the Foot Skating Ratio (FSR), which measures the proportion of such frames.

Besides, to evaluate the deterministic results, we use two metrics to measure the quality of the generated results.

- **LVD:** Landmark Velocity Difference evaluates the speed difference between the ground-truth and generated facial landmarks, assessing the alignment between the spoken input and the corresponding face motion.
- **LD:** Landmark Distance is used to measure the difference between the ground-truth and generated facial landmarks, including jaw joints and lip shape. In our evaluation, we adopt L2 distance as the measurement.

4.2 Comparison Results

Quantative Evaluation. For body and hand motion generation, Table 1 presents the quantitative comparison on BEAT2 dataset with several baselines. It can be seen that our method outperforms other methods in most metrics. Specifically, our method significantly improves the performance on FID (both FID-k and FID-g), which indicates that our model

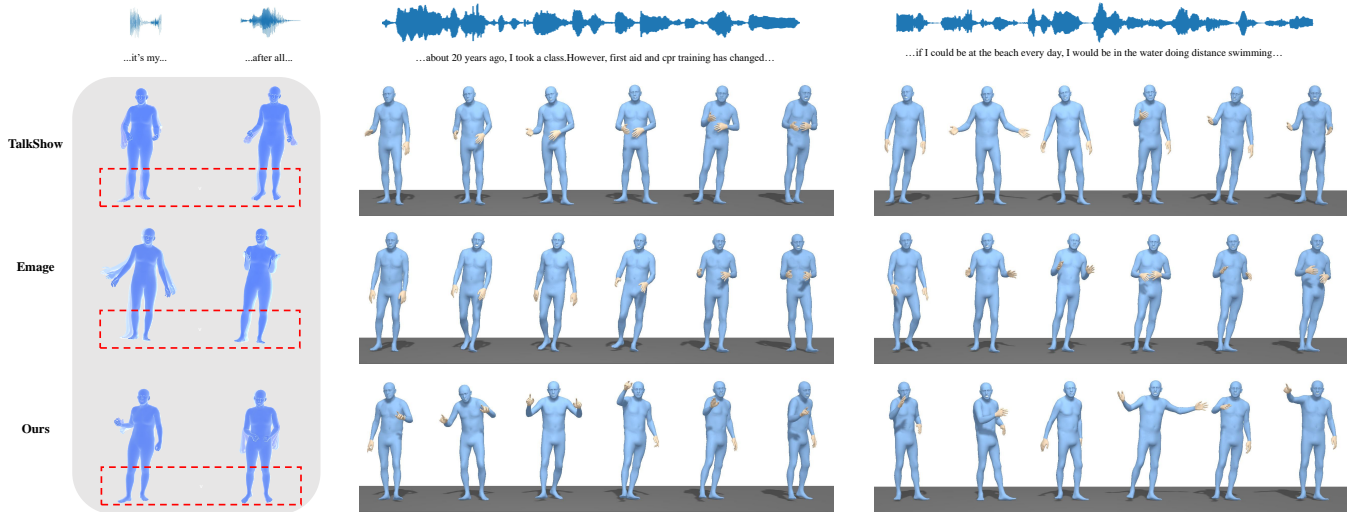


Fig. 5: Qualitative results compared with TalkShow [1] and EMAGE [46]. The left subfigure shows the continuity and the smoothness of the generated motions, and the right subfigure presents the diversity of the results. In the left subfigure, the first row shows the two different audio inputs, the second row presents the related text, and the other rows show the generated results by different methods. Each sample consists of five frames extracted at intervals of 2/15 seconds from a generated motion clip. Lighter colors represent past frames.

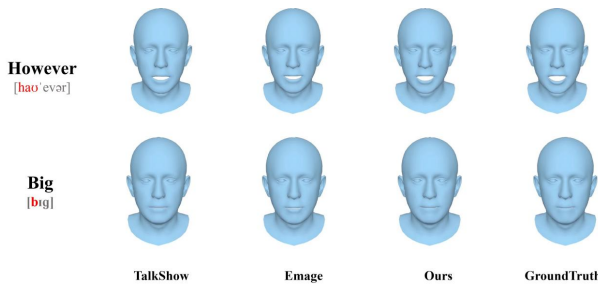


Fig. 6: Comparison of generated facial movements.

generates more realistic body and hand motion sequences. The metric FID-k focuses on the speed and acceleration of movement, reflecting the physical realism of the generated motions. Results generated by previous often contain foot-skating frames and disharmony contacts between ground and feet, which causes unnatural speed and acceleration of the movements. While our method can generate more natural and smooth motion sequences due to our hybrid representation, which leads to the high improvements on FID-K. Besides, the comparison on FSR demonstrates that our model can generate more smooth motions and can alleviate foot skating in generated motions, which suggests the effectiveness of our hybrid representation. Also, our method achieves the best performance on BeatAlign score. This denotes that the motion sequences generated by our method are more rhythm-consistent with the input audio. Moreover, the diversity of our results outperforms other baselines, which indicates our contrastive motion learning method can generate more distinctive motion representations. The diffusion-based models, *i.e.*, DiffStyle, requires a large number of diffusion steps, leading to 7 times longer than our method to generate a 4-second motion sequence (5.80s vs. 0.78s). The detailed comparison on inference time

can be found in the supplemental document.

Table 2 shows the quantitative comparison on BEAT2 dataset for face motion generation. According to the comparison, our face generator can synthesize more accurate results aligned with the input audio. This proves the superior performance of our face generator.

Qualitative Evaluation. Fig. 5 shows the qualitative comparison compared with TalkShow [1] and EMAGE [46]. In the left subfigure, each sample consists of five frames extracted at intervals of 2/15 seconds from a generated motion clip. It can be seen that the foot movements generated by TalkShow and EMAGE change rapidly in a short time, leading to foot skating. In contrast, due to our hybrid point representation, our model can generate continuous and smooth motion movements. Besides, our method generates realistic movements corresponding to the audio input, *e.g.*, hands down when emphasizing “after all”. The right subfigure in Fig. 5 presents two generated results given two audio inputs. It can be seen that our method can generate more diverse body and hand motions for a single sentence, which benefits from our contrastive motion learning method. Fig. 6 presents the comparison of generated facial movements. The results show that our method can generate more accurate motions for both open-mouth and closed-mouth sounds. More dynamic results and analyses of different methods can be found in the supplementary material. We also provide some long generation samples (over 3 minutes) in the supplemental video.

To boost the diversity of the generated results, we design a contrastive motion learning method according to an intuitive idea that the generated motion of specific audio should be different from the generated motions of other speakers and other audios. To verify the performance of our model, we visualize generated results using tSNE method to present the diversity of different speakers with different speech audios and the diversity of a single speaker with

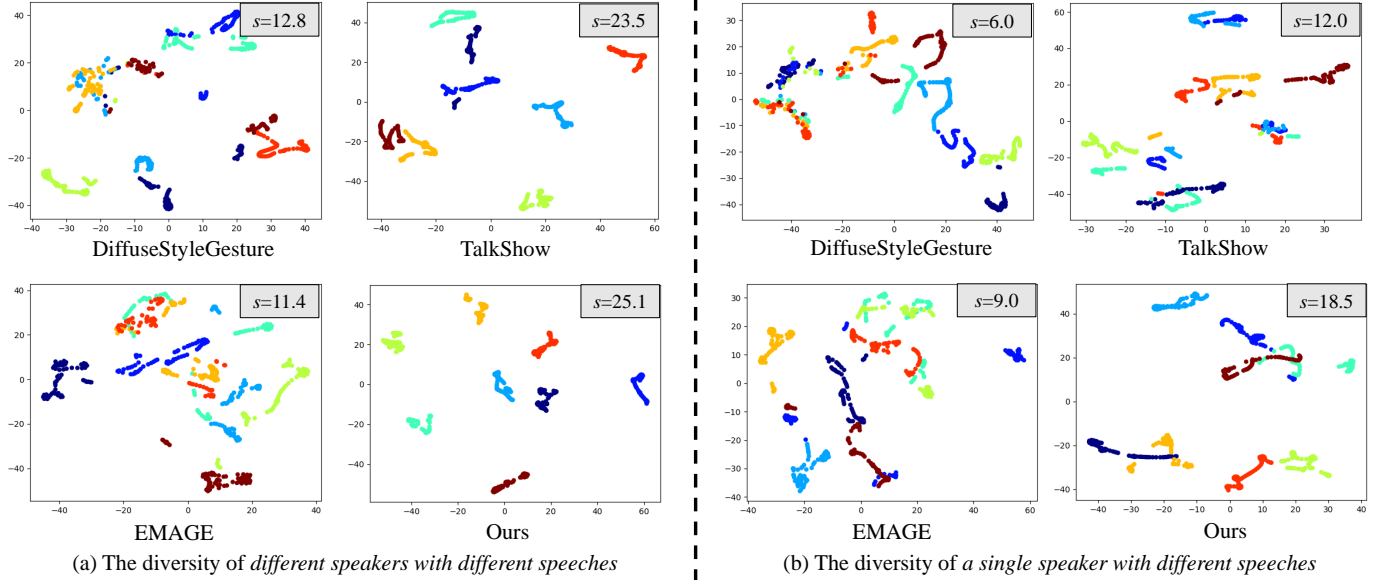


Fig. 7: Visualization of the diversity of different speakers with different speech audios and the diversity of a single speaker with different speech audios compared with three state-of-the-art methods. The inter-class distance of each method is shown in the top right corner of each figure. Different colors in (a) indicate the visualization of the generated results conditioned on different speakers, and each speaker in the same method has a specific speech audio. Different colors in (b) show the visualization of different speech audios talked by the same speaker.

TABLE 3: Ablation studies on BEAT2 dataset for different motion representations. Note that we report $LD \times 10^2$ for vertices.

Representation	Keypoints		Vertices		Param.
	LVD ↓	LD ↓	LVD ↓	LD ↓	
3D (TalkShow)	0.4409	14.42	26.71	10.87	82.4 M
6D	0.4514	15.76	26.98	11.86	82.7 M
Surface Points	0.3095	10.85	21.68	7.896	88.5 M
Keypoints	0.2836	10.68	19.39	8.595	84.0 M
Ours	0.2744	7.936	16.68	6.050	89.4 M

different speech audios. Fig. 7 illustrates the visualization of the generated results. We also give the inter-class distance (s) to present the diversity between the different clusters. The inter-class distance refers to the measure of separation between different groups. Fig. 7 (a) is obtained from 8 speakers with different speech audios, which indicates that the generated motions should be different. It can be seen that our model has better performance, which can generate more distinctive motions different from other speakers. On the contrary, the other methods struggle to generate distinctive motions, leading to similar motions even among different speakers with different sentences. This is also the reason that the inter-class distance of TalkShow (23.5) is closer to our method (25.1). Fig. 7 (b) illustrates the diversity of a single speaker with different speech audios. We can see that with different speech audios, our method achieves better cluster performance and gets the best inter-class distance, which indicates that our method can generate more diverse results for a single speaker.

4.3 Ablation Study

We conduct extensive experiments to validate the effectiveness of our representation and our model.

Effectiveness of hybrid point representation. To validate the effectiveness of our hybrid point representation, instead of conducting experiments on the motion generation task, we compare different representations on the motion reconstruction task. Specifically, we replace our representation with different representations for the first stage of our body generator, *i.e.*, quantized motion modeling (Sec. 3.1.2). We employ the 3D representation, 6D representation, surface points of SMPL-X body mesh, and keypoints of SMPL-X body mesh. Note that aside from the dimensions of the input and output layers, all other layer configurations, such as the number of layers and hidden dimensions, are identical. It should be noted that the model with 3D representation has the same representation as TalkShow.

We adopt landmark velocity difference (LVD) and landmark distance (LD) to evaluate the quality of reconstructed motion. To evaluate different representations, we extract keypoints and vertices of reconstructed SMPL-X body meshes to compute LVD and LD metrics. In addition, we report the number of parameters (Param.) for different representations. Table 3 presents the quantitative results on BEAT2 dataset for different motion representations. It can be seen that compared with the surface point representation, the parametric representation can generate inaccurate results (high LD score) with jitters (high LVD score). It is interesting that the performance of 6D representation is lower than 3D representation. The main reason is that speech motion, involves a lot of local rotations, which typically involve small angle changes. Therefore, the 3D axis-angle representation might more easily capture these local characteristics, leading to better performance in the re-

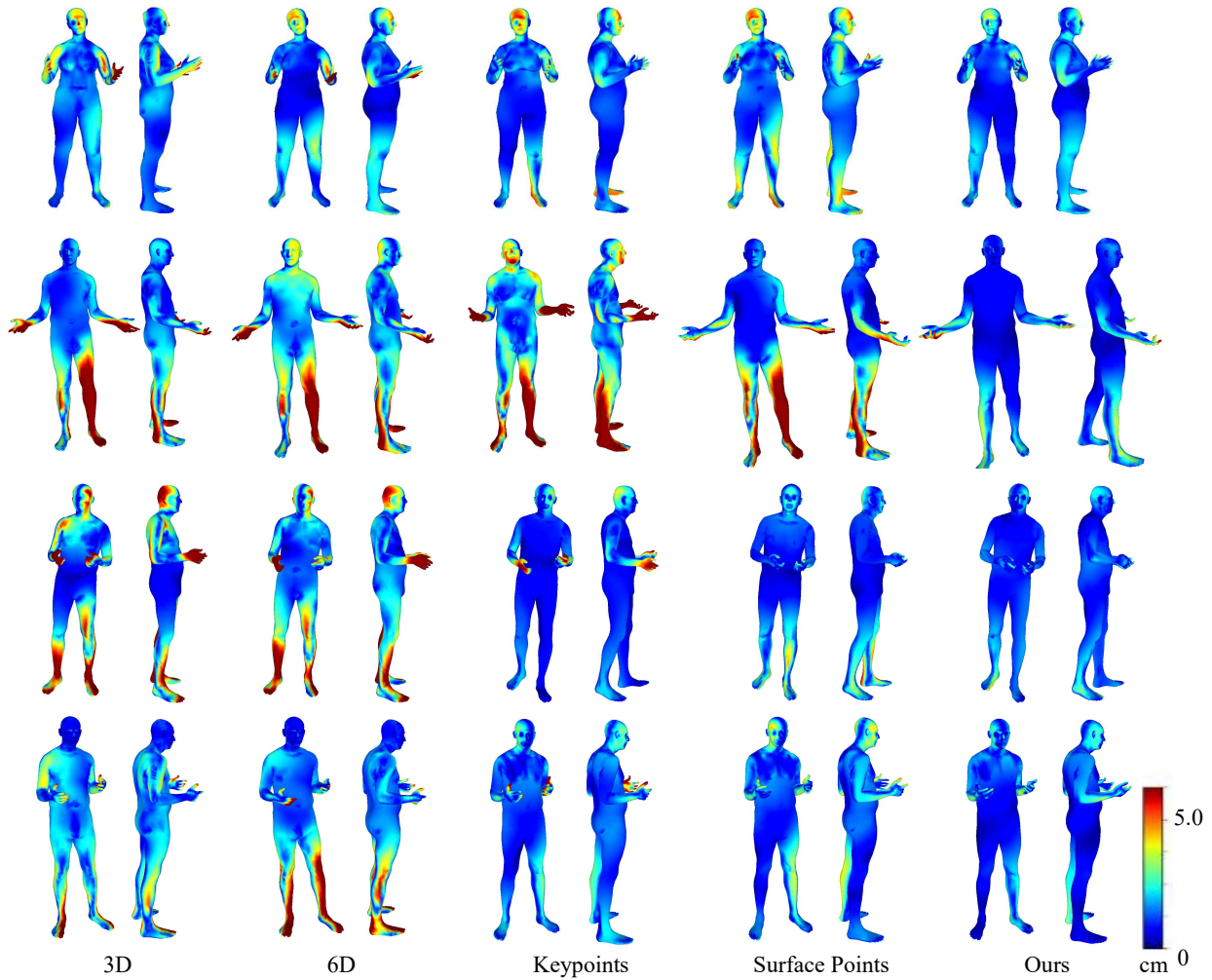


Fig. 8: Visualization of reconstructed motion errors. The errors between the reconstructed motions and the ground-truths are color-coded on the reconstructed models for visual inspection. Each row shows reconstructed error for each case.

construction task. Comparing surface points representation and keypoints representation, our hybrid representation can achieve more accurate and smoother results, which indicates that our representation is suitable as a motion representation. This is also the reason that our model achieves better results in Table 1 compared with other methods.

Fig. 8 shows the qualitative results. It can be seen that the models based on 3D and 6D representations can produce errors for local body parts, such as hands and legs. The keypoint-based model has better performance, however, the arms are difficult to recover due to a lack of global constraints of the pose rotation. The point-based model can recover more accurate arms, but it struggles to reconstruct the head and legs accurately due to redundant information for some body parts. Compared with these traditional representations, our hybrid point representation can recover the body accurately, which indicates its strong capacity for motion reconstruction. With this powerful representation, we can learn a good motion codebook to further help generate natural motions from speech audios.

Effectiveness of contrastive motion learning. We introduce a novel contrastive motion learning to boost the diversity of generated results. To validate the effectiveness of this method, we train a body generator without the contrastive

TABLE 4: Ablation studies on BEAT2 dataset for contrastive motion learning.

Model	FID-k ↓	FID-g ↓	BeatAlign ↑	Div-in ↑	Div-out ↑
w/o CM	3.865	5.972	0.5767	0.8511	2.122
Full	3.839	4.359	0.5603	0.9700	2.264

motion learning method (w/o CM) as a comparison. The other settings are the same as our full model.

Table 4 presents the quantitative comparison of BEAT2 dataset. By comparing the model without the contrastive motion learning method and our full model, it can be seen that the contrastive motion learning method can improve the diversity of generated results. It is worth noting that the quality-related scores of the model without the contrastive motion learning method still outperform the scores of Talk-Show and Audio2Gesture, which suggests the effectiveness of our hybrid point representation. However, the BeatAlign score of our full model is slightly lower than that of the model without the contrastive motion learning method, which is possibly due to the increase in diversity.

Fig. 9 shows the visualization of motions generated by our model. Compared with w/o CM and our full model, it

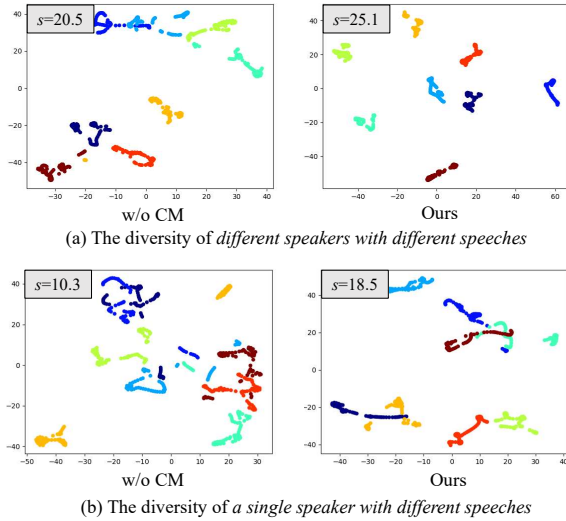


Fig. 9: Visualization of the diversity of different speakers with different speech audios (a) and the diversity of a single speaker with different speech audios (b). The inter-class distance is shown at the top left corner of each figure.

TABLE 5: Ablation studies on BEAT2 dataset for different numbers of codebooks. Note that we report $LD \times 10^2$ for vertices.

Number	Keypoints		Vertices	
	LVD ↓	LD ↓	LVD ↓	LD ↓
1	0.2776	8.098	16.78	6.226
2	0.2683	8.002	16.71	6.179
3	0.2744	7.936	16.68	6.050

can be seen that the contrastive learning method can help the model to generate more distinctive motions for different speakers and different speech audios. The increase in the inter-class distances of our full model also verifies this.

Effectiveness of separate codebooks. We use three codebooks, *i.e.*, the body codebook, the left-hand codebook, and the right-hand codebook, to model more detailed motions for the body and hands, respectively. To validate the effectiveness of this design, we conduct experiments using different numbers of codebooks. Specifically, we train motion reconstruction models using one codebook, two codebooks (hand and body codebooks), and three codebooks (our current design). Table 5 illustrates the quantitative results. As shown, increasing the number of codebooks improves the performance of motion reconstruction. Although the landmark velocity distance (LVD) for joints is better with two codebooks, the other metrics are lower compared to the model with three codebooks. Therefore, we adopt three separate codebooks in our model.

4.4 User Study

To better evaluate the proposed method qualitatively, we conduct user studies to analyze the performance of generated motions. Our questionnaire consists of 10 cases, with each case comprising three questions. The motions for each case in the video are generated based on the same audio

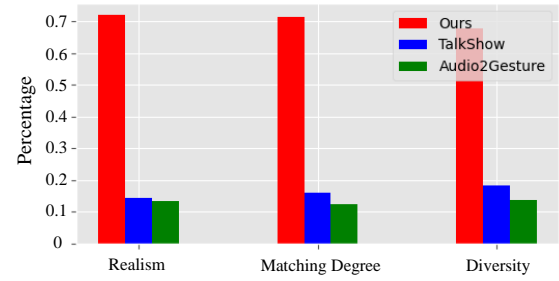


Fig. 10: The percentage of each method considered to be ranked first in different perspectives.

input using Audio2Gesture [36], TalkShow [1], and our method. Users are required to rank the results of different methods from the following three perspectives: 1) the realism of the generated motion; 2) the matching degree between the generated motion and the input audio; and 3) the diversity of the generated motion. We have collected answers from 153 participants, including 85 males and 68 females with different ages (3 users below 18, 105 users between 18 and 40, 44 users between 40 and 60, and 1 user beyond 60).

We evaluate the percentage of each method considered to be ranked first in different perspectives. Fig. 10 shows the statistical results. It is evident that more than 70% of users believe that our method generates more realistic results, which indicates that our hybrid point representation can produce more smooth and natural results. Besides, compared to other methods, more than 70% of users think that our method produces results more in line with the audio rhythm, and 68% of users believe that our method can generate more diverse motion. We also conducted a significance test between our method and the other two methods, which indicates that our method shows statistically significant differences compared to the other two methods. These results demonstrate that our method surpasses other methods in terms of diversity and alignment with audio signals, which suggests the effectiveness of our hybrid point representation and our contrastive motion learning method. Details of the user study and the significance test can be found in the supplemental document.

4.5 Generalization

Our model generates natural and diverse motions from only audio signals. Therefore, even with training with English data of BEAT2 dataset [46], our model can be generalized to other languages. Fig. 11 shows the qualitative results. We can see that with the specific rhythm, our model can generate corresponding motions.

4.6 Application

Thanks to our hybrid point representation, our model can output SMPLX parameters to support many applications. In this section, we obtain an avatar reconstructed from [69], and animate it using the outputs of our model. By doing so, we can get animated avatars directly using speech, which can be used in VR/AR, virtual live, and human-computer interaction. Since the expression basis of the facial model, *i.e.*, ARKit, used in AvatarRex differs from ours, *i.e.*, SMPLX, we retrained our model using the expression format of ARKit [70]. As a result, with the output motions

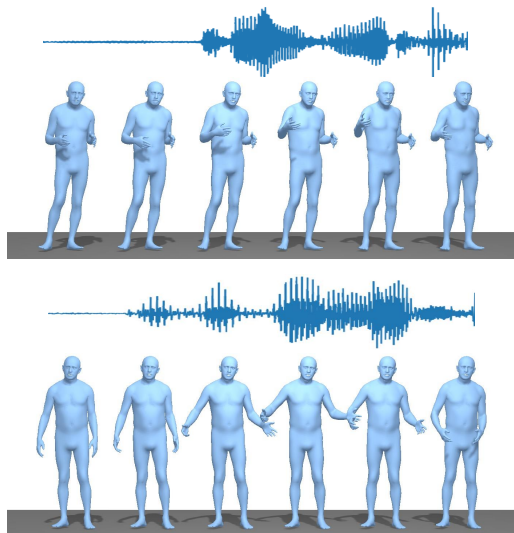


Fig. 11: Visualization of generalization across different languages and sentences. The first sequence is in Chinese, and the second sequence is in Spanish.



Fig. 12: Avatar animation results. Given a speech audio, the generated motion can be used to animate reconstructed avatars.

generated by our method, we can directly animate the avatar reconstructed using AvatarRex. To showcase additional results, we adopt the recent avatar reconstruction method ExAvatar [71] to produce animation results based on our generated motion sequence. Notably, we can directly use the first 50 dimensions of our expression parameters as the expression parameters for ExAvatar, which ensures compatibility with our method without the need for retraining a new model. Fig. 12 presents the avatar animation results reconstructed from AvatarRex (the first row) and ExAvatar (the second row). It can be seen that the generated motion can be applied to avatar animation, and can obtain promising results. The dynamic results can be found in the supplementary video.

5 CONCLUSIONS AND DISCUSSIONS

Conclusion. In this paper, we propose a novel framework, named SpeechAct, with a face generator and a body generator for whole-body motion generation from speech audio. The face generator is used to generate deterministic results

using an encode-decoder architecture. For the body generator, we propose a hybrid point representation for body and hand to constrain the global surface and capture local details for 3D body, which can achieve accurate yet continuous results. Based on our representation, we design a two-stage model with a novel contrastive motion learning method to achieve diverse body and hand motions. Experimental results demonstrate that our model can generate natural and diverse human motion. We give some promising results for generalization across other languages. Moreover, an application of speech-driven avatar animation is given to show the potential of VR/AR and human-computer interaction.

Limitation and Future Work. To achieve natural results, we estimate the global position using a residual estimation manner. Specifically, we estimate position residuals and add previous residuals to obtain the current position, which can help to generate smooth results. However, for long motion generation, the position error will be added and lead to a sliding effect for some specific cases. The visualization and discussion of some samples can be found in the supplemental document. In the future, we will consider more temporal constraints to generate more smooth results. Besides, our model can generate diverse results with distinctive styles. However, we only use the speaker embedding as the style condition, which limits the effectiveness of controllability. In the future, we will add more control inputs, e.g., text, as conditional inputs, and leverage the speaker information to achieve more controllable human motion synthesis.

ACKNOWLEDGEMENTS

This work was supported in part by National Key R&D Program of China (2023YFC3082100), National Natural Science Foundation of China (62122058 and 62171317), and Science Fund for Distinguished Young Scholars of Tianjin (No. 22JCJQC00040). We are grateful to Associate Editor and anonymous reviewers for their help in improving this paper.

REFERENCES

- [1] H. Yi, H. Liang, Y. Liu, Q. Cao, Y. Wen, T. Bolkart, D. Tao, and M. J. Black, "Generating holistic 3d human motion from speech," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.
- [2] S. Yang, Z. Wu, M. Li, Z. Zhang, L. Hao, W. Bao, M. Cheng, and L. Xiao, "DiffuseStyleGesture: stylized audio-driven co-speech gesture generation with diffusion models," in *Proc. IJCAI*, 2023, pp. 5860–5868.
- [3] I. Habibi, W. Xu, D. Mehta, L. Liu, H.-P. Seidel, G. Pons-Moll, M. Elgharib, and C. Theobalt, "Learning speech-driven 3d conversational gestures from video," in *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, 2021.
- [4] B. H. Le, X. Ma, and Z. Deng, "Live speech driven head-and-eye motion generators," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 11, pp. 1902–1914, 2012.
- [5] A. Aristidou, A. Yiannakidis, K. Aberman, D. Cohen-Or, A. Shamir, and Y. Chrysanthou, "Rhythm is a dancer: Music-driven motion synthesis with global structure," *IEEE Trans. Vis. Comput. Graph.*, 2022.
- [6] J. Li, D. Kang, W. Pei, X. Zhe, Y. Zhang, L. Bao, and Z. He, "Audio2gestures: Generating diverse gestures from audio," *IEEE Trans. Vis. Comput. Graph.*, 2023.
- [7] Z. Wang, J. Chai, and S. Xia, "Combining recurrent neural networks and adversarial training for human motion synthesis and control," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 1, pp. 14–28, 2019.
- [8] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik, "Learning individual styles of conversational gesture," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

- [9] T. Ao, Q. Gao, Y. Lou, B. Chen, and L. Liu, "Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings," *ACM Trans. Graph.*, vol. 41, no. 6, pp. 1–19, 2022.
- [10] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, "Generating diverse and natural 3d human motions from text," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 5152–5161.
- [11] N. Kolotouros, G. Pavlakos, and K. Daniilidis, "Convolutional mesh regression for single-image human shape reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.
- [12] Y. Tian, H. Zhang, Y. Liu, and L. Wang, "Recovering 3d human mesh from monocular images: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.
- [13] Y. Zhang, M. J. Black, and S. Tang, "We are more than our joints: Predicting how 3d bodies move," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.
- [14] X. Li, J. Zhang, Y.-K. Lai, J. Yang, and K. Li, "High-quality animatable dynamic garment reconstruction from monocular videos," *IEEE Trans. Circuit Syst. Video Technol.*, 2023.
- [15] H. Zhao, J. Zhang, Y.-K. Lai, Z. Zheng, Y. Xie, Y. Liu, and K. Li, "High-fidelity human avatars from a single rgb camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 15904–15913.
- [16] B. Degardin, J. Neves, V. Lopes, J. Brito, E. Yaghoubi, and H. Proença, "Generative adversarial graph convolutional networks for human action synthesis," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1150–1159.
- [17] Z. Cao, H. Gao, K. Mangalam, Q.-Z. Cai, M. Vo, and J. Malik, "Long-term human motion prediction with scene context," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 387–404.
- [18] X. Ma, J. Su, C. Wang, W. Zhu, and Y. Wang, "3D human mesh estimation from virtual markers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 534–543.
- [19] X. Zhang, B. L. Bhatnagar, S. Starke, V. Guzun, and G. Pons-Moll, "Couch: Towards controllable human-chair interactions," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2022, pp. 518–535.
- [20] M. Hassan, D. Ceylan, R. Villegas, J. Saito, J. Yang, Y. Zhou, and M. J. Black, "Stochastic scene-aware motion prediction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 11374–11384.
- [21] K. Li, Y. Liu, Y.-K. Lai, and J. Yang, "Mili: Multi-person inference from a low-resolution image," *Fundamental Research*, vol. 3, no. 3, pp. 434–441, 2023.
- [22] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3D speaking styles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.
- [23] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "Faceformer: Speech-driven 3d facial animation with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.
- [24] J. Cassell, H. H. Vilhjálmsón, and T. Bickmore, "Beat: the behavior expression animation toolkit," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. New York, NY, USA: Association for Computing Machinery, 2001, p. 477–486.
- [25] S. Kopp and I. Wachsmuth, "Synthesizing multimodal utterances for conversational agents," *Computer animation and virtual worlds*, vol. 15, no. 1, pp. 39–52, 2004.
- [26] S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun, "Gesture controllers," in *Acm siggraph*, 2010.
- [27] K. Takeuchi, S. Kubota, K. Suzuki, D. Hasegawa, and H. Sakuta, "Creating a gesture-speech dataset for speech-based automatic gesture generation," in *the SIGCHI Conference on Human Factors in Computing Systems*, 2017.
- [28] Y. Yoon, W.-R. Ko, M. Jang, J. Lee, J. Kim, and G. 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)*. IEEE, 2019.
- [29] C. Ahuja, D. W. Lee, R. Ishii, and L.-P. Morency, "No gestures left behind: Learning relationships between spoken language and freeform gestures," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, 2020, pp. 1884–1895.
- [30] U. Bhattacharya, E. Childs, N. Rewkowski, and D. Manocha, "Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning," in *ACM Int. Conf. Multimedia*. New York, NY, USA: Association for Computing Machinery, 2021, p. 2027–2036.
- [31] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström, "Analyzing input and output representations for speech-driven gesture generation," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 2019.
- [32] T. Kucherenko, P. Jonell, S. Van Waveren, G. E. Henter, S. Alexandersson, I. Leite, and H. Kjellström, "Gesticulator: A framework for semantically-aware speech-driven gesture generation," in *Proceedings of the 2020 international conference on multimodal interaction*. New York, NY, USA: Association for Computing Machinery, 2020, p. 242–250.
- [33] C. Ahuja, D. W. Lee, Y. I. Nakano, and L.-P. Morency, "Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Heidelberg: Springer-Verlag, 2020, p. 248–265.
- [34] X. Liu, Q. Wu, H. Zhou, Y. Xu, R. Qian, X. Lin, X. Zhou, W. Wu, B. Dai, and B. Zhou, "Learning hierarchical cross-modal association for co-speech gesture generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.
- [35] Y. Yoon, B. Cha, J.-H. Lee, M. Jang, J. Lee, J. Kim, and G. Lee, "Speech gesture generation from the trimodal context of text, audio, and speaker identity," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 1–16, 2020.
- [36] J. Li, D. Kang, W. Pei, X. Zhe, Y. Zhang, Z. He, and L. Bao, "Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.
- [37] S. Qian, Z. Tu, Y. Zhi, W. Liu, and S. Gao, "Speech drives templates: Co-speech gesture synthesis with learned templates," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.
- [38] P. J. Yazdian, M. Chen, and A. Lim, "Gesture2Vec: Clustering gestures using representation learning methods for co-speech gesture generation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022.
- [39] S. Alexanderson, G. E. Henter, T. Kucherenko, and J. Beskow, "Style-controllable speech-driven gesture synthesis using normalising flows," *Comput. Graph. Forum*, vol. 39, no. 2, pp. 487–496, 2020.
- [40] Z. Yang, Y.-H. Wen, S.-Y. Chen, X. Liu, Y. Gao, Y.-J. Liu, L. Gao, and H. Fu, "Keyframe control of music-driven 3d dance generation," *IEEE Trans. Vis. Comput. Graph.*, 2023.
- [41] J. Ho, A. Jain, and P. Abbeel, "Denosing diffusion probabilistic models," *Proc. Adv. Neural Inform. Process. Syst.*, vol. 33, pp. 6840–6851, 2020.
- [42] T. Ao, Z. Zhang, and L. Liu, "Gesturediffuclip: Gesture diffusion model with clip latents," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–18, 2023.
- [43] S. Alexanderson, R. Nagy, J. Beskow, and G. E. Henter, "Listen, denoise, action! audio-driven motion synthesis with diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–20, 2023.
- [44] R. Li, Y. Zhang, Y. Zhang, H. Zhang, J. Guo, Y. Zhang, Y. Liu, and X. Li, "Lodge: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives," *arXiv preprint arXiv:2403.10518*, 2024.
- [45] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik, "Learning individual styles of conversational gesture," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.
- [46] H. Liu, Z. Zhu, G. Becherini, Y. Peng, M. Su, Y. Zhou, N. Iwamoto, B. Zheng, and M. J. Black, "Emage: Towards unified holistic co-speech gesture generation via masked audio gesture modeling," *arXiv preprint arXiv:2401.00374*, 2023.
- [47] S. Mahendran, H. Ali, and R. Vidal, "A mixed classification-regression framework for 3d pose estimation from 2d images," in *Proc. Brit. Mach. Vis. Conf.*, 2018.
- [48] C. Jiang, F. Gao, B. Ma, Y. Lin, N. Wang, and G. Xu, "Masked and adaptive transformer for exemplar based image translation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.
- [49] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 319–345.
- [50] D. d. S. Edirimuni, X. Lu, G. Li, and A. Robles-Kelly, "Contrastive learning for joint normal estimation and point cloud filtering," *IEEE Trans. Vis. Comput. Graph.*, pp. 1–15, 2023.
- [51] A. Sanghi, "Info3D: Representation learning on 3d objects using mutual information maximization and contrastive learning," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 626–642.

- [52] Z. Zhao, J. Bai, D. Chen, D. Wang, and Y. Pan, "Taming diffusion models for music-driven conducting motion generation," in *Proc. AAAI*, vol. 1, no. 1, 2023, pp. 40–44.
- [53] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, "Generating 3D faces using convolutional mesh autoencoders," in *Proc. Eur. Conf. Comput. Vis.*, 2018.
- [54] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.
- [55] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," in *Proc. Adv. Neural Inform. Process. Syst.*, 2017.
- [56] L. Siyao, W. Yu, T. Gu, C. Lin, Q. Wang, C. Qian, C. C. Loy, and Z. Liu, "Bailando: 3d dance generation by actor-critic gpt with choreographic memory," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.
- [57] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 15745–15753.
- [58] Y. Sun, L. Dong, S. Huang, S. Ma, Y. Xia, J. Xue, J. Wang, and F. Wei, "Retentive network: A successor to transformer for large language models," *arXiv preprint arXiv:2307.08621*, 2023.
- [59] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [60] Z. Ma, X. Zhu, G. Qi, C. Qian, Z. Zhang, and Z. Lei, "Diffspeaker: Speech-driven 3d facial animation with diffusion transformer," *arXiv preprint arXiv:2402.05712*, 2024.
- [61] Q. Zhao, P. Long, Q. Zhang, D. Qin, H. Liang, L. Zhang, Y. Zhang, J. Yu, and L. Xu, "Media2face: Co-speech facial animation generation with multi-modality guidance," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–13.
- [62] Z. Sun, T. Lv, S. Ye, M. Lin, J. Sheng, Y.-H. Wen, M. Yu, and Y.-j. Liu, "Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 43, no. 4, pp. 1–9, 2024.
- [63] S. Stan, K. I. Haque, and Z. Yumak, "Facediffuser: Speech-driven 3d facial animation synthesis using diffusion," in *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games*, 2023, pp. 1–11.
- [64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [65] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," *Proc. Adv. Neural Inform. Process. Syst.*, 2017.
- [66] K. Onuma, C. Faloutsos, and J. K. Hodgins, "FMDistance: A fast and effective distance function for motion capture data," in *Eurographics (Short Papers)*, 2008.
- [67] E. Ng, H. Joo, L. Hu, H. Li, T. Darrell, A. Kanazawa, and S. Ginosar, "Learning to listen: Modeling non-deterministic dyadic facial motion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 20395–20405.
- [68] K. Karunratanakul, K. Preechakul, S. Suwajanakorn, and S. Tang, "Guided motion diffusion for controllable human motion synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 2151–2162.
- [69] Z. Zheng, X. Zhao, H. Zhang, B. Liu, and Y. Liu, "Avatarrex: Real-time expressive full-body avatars," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–19, 2023.
- [70] J. Linowes and K. Babilinski, *Augmented reality for developers: Build practical augmented reality applications with unity*, ARCore, ARKit, and Vuforia, 2017.
- [71] G. Moon, T. Shiratori, and S. Saito, "Expressive whole-body 3d gaussian avatar," in *Proc. Eur. Conf. Comput. Vis.*, 2025, pp. 19–35.



Jinsong Zhang received B.E. and M.E. degrees from Tianjin University, China, in 2018 and 2021, respectively, where he is currently pursuing a Ph.D. degree in computer science. His research interests mainly include computer vision and computer graphics.



Minjie Zhu received her bachelor's degree from Anhui University in 2018 and she is currently studying for master's degree in Tianjin University. Her main research direction is computer vision.



Yuxiang Zhang received the B.E. degree from the Tsinghua University, Beijing, China, in 2019, and the Ph.D. degree from the Tsinghua University, Beijing, China, in 2024. His research interests include 3D vision, human motion capture, 3D human reconstruction and close human interaction recovery.



Zerong Zheng received the B.S. degree and the PhD degree in Department of Automation, Tsinghua University in July 2018 and July 2023, respectively. He is currently a algorithm scientist in Bytedance. His current research interests include computer vision and computer graphics.



Yebin Liu (Member, IEEE) received the B.E. degree from the Beijing University of Posts and Telecommunications, China, in 2002 and the Ph.D. degree from the Automation Department, Tsinghua University, Beijing, China, in 2009. He is currently a full professor with Tsinghua University. He was a research fellow in the Computer Graphics Group of the MaxPlanck Institute for Information, Germany, in 2010. His research areas include computer vision, computer graphics, and computational photography.



Kun Li (Senior Member, IEEE) received the B.E. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2006, and the master and Ph.D. degrees from Tsinghua University, Beijing, in 2011. She is currently a Professor with the College of Intelligence and Computing, Tianjin University, Tianjin, China. She was the recipient of the CSIG Shi Qingyun Award for Women Scientists. Her research interests include 3D reconstruction and AIGC.