Supplemental Document: Learning Hierarchical Cross-Modal Association for Co-Speech Gesture Generation

Xian Liu[1], Qianyi Wu[2], Hang Zhou[1], Yinghao Xu[1], Rui Qian[1], Xinyi Lin[3],
Xiaowei Zhou[3], Wayne Wu[4], Bo Dai[5], Bolei Zhou[1]
[1]The Chinese University of Hong Kong [2]Monash University [3]Zhejiang University
[4]SenseTime Research [5]S-Lab, Nanyang Technological University
{alvinliu@ie, zhouhang@link, xy119@ie, qr021@ie, bzhou@ie}.cuhk.edu.hk, qianyi.wu@monash.edu,
{shinylin, xwzhou}@zju.edu.cn, wuwenyan@sensetime.com, bo.dai@ntu.edu.sg

1. More Details about Dataset

Choice of Data and Data Collection. Many learning-based approaches use motion and gesture training data captured in a MoCap studio with complex motion capture systems [10, 12]. They can acquire more accurate human motion data compared to automatic annotations on internet videos. However, such methods have the following drawbacks: 1) Owing to the high cost of MoCap data, it is hard to build a large-scale corpus of data covering various speaking contents and styles. For example, the length of MSP-AVATAR [15] and Personality Dyads Corpus [17] are less than 3h. 2) When capturing co-speech gesture data in the studio, the actors/actresses are asked to deliberately talk with their arms and hands moving, which contributes to the unnaturalness and exaggeration of captured motion data. Therefore, we follow the previous works [6, 18, 19] to collect internet videos and annotate 3D human pose as pseudo ground truth for later training. Specifically, Ginosar *et al.* [6] and Habibie *et al.* [7] use a speaker-specific gesture dataset of a very small number of speakers, *i.e.*, 10 speakers in [6] and 6 speakers in [7], making them unable to transfer to general speaking styles. TED Gesture dataset is proposed by Yoon *et al.* [18] which contains over 1,700 TED talks covering diverse topics and speaker styles. Following Yoon *et al.* [18], we propose to build our TED-Expressive dataset based on the raw videos of TED talks. Differently, since the flexible finger movement matters a lot when people talk, we add the information of finger keypoints for more expressive co-speech gesture dataset establishment. We collect internet videos from the official TED channel on YouTube.[1] We finally get 1,764 videos and their corresponding text transcripts.

[1]We obey the TED Talks Team's Creative Commons License (CC BY-NC-ND 4.0 International) by referencing all the video links shown in our papers. We sincerely thank the permission of the TED Talks Team for using the videos, audios and transcripts in this paper.

Pose Annotation and Post-Processing. To get the reliable pseudo ground truth of co-speech human upper body pose with finger keypoints, we leverage the state-of-art 3D human motion estimator ExPose [4] for annotation. Similar to the step of [19], we segment videos into smaller shots by their scenes and annotate the 2D human pose of each frame by OpenPose [3]. With the 2D pose prior provided by OpenPose, we use ExPose [4] to annotate 3D upper body keypoints. Concretely, we use 43 keypoints, including 13 upper body joints (spine, neck, nose, left/right eyes, ears, shoulders, elbows and wrists, totally $13 = 3+5*2$) and 30 finger joints (3 joints for each finger, totally $30 = 3*5*2$). Then we select shots of interest under the following conditions: 1) the above mentioned 43 keypoints of speaker are visible for more than 50% frames of a clip; 2) the speaker should not remain almost still in the whole shot, *i.e.*, the variance of motion is quite small; 3) the clip is longer than 5s. The statistics of TED Gesture and TED-Expressive dataset are recorded in Table 1. For the TED Gesture dataset, we randomly split the segments into the 80% training set, 10% validation set, 10% test set and finally get 199,384; 26,795; and 25,930 segments in each partition.

Pose Representation and Quality. After the filtering process, we effectively eliminate the influence of bone length by normalizing them into 42 unit directional vectors to represent each bone. Such 3D representation is invariant to root joint motion and body shape, thus making it more stable in the training phase. At the inference stage, the mean bone length over the training set is multiplied to predicted bone vectors for visualized results. The whole pipeline is automated, which facilitates us to build a large corpus of co-speech gesture dataset. Figure 3 shows the correspondence between keypoint index and joints. We can see that there are totally 43 annotated upper body keypoints, which are then transformed into 42 unit direction vectors as mentioned above.

Statistics	# of Videos	Interest Shots Length	# of Segments	Interest Ratio
TED Gesture [18, 19]	1,766	106.1h	252,109	25%
TED-Expressive	1,764	100.8h	240,447	21%

Table 1. Statistics of the TED Gesture and TED-Expressive dataset.

As the pose annotations serve as pseudo ground truth in our pipeline, the quality of annotations is crucial for training. However, since the pose representation is 3D, we can not follow Ginosar *et al.* [6] to evaluate the quality of annotations by automatic pipeline against human annotations. But the high performance of ExPose on benchmark datasets and our filtering algorithm guarantee that the data quality is good enough for utilization. Please refer to ExPose [4] for the detailed quantitative 3D pose estimation results on benchmark dataset. Overall, we use the open-source code of Trimodal [18], OpenPose [3] and ExPose [4] following their licenses[2].

Speech Audio Pre-Processing. The speech audios accompanied TED videos are raw waveforms, which are processed to 16kHZ and convert to mel-spectrograms as 2D time-frequency representations for more compact information preservation. The FFT window size is 1024 and the hop length is 512.

Speech Text Pre-Processing. We collect speech text input with the transcripts of TED videos. Then, the onset timestamps of each word are extracted by the Gentle forced aligner [14] to insert padding tokens. For example, for the speech text "Good morning everyone", if there is a short pause between the word "morning" and "everyone", then the padded word sequence is "Good morning $\diamond$ $\diamond$ everyone" as padded by Gentle if the time of this sentence is 5. Following the process of [18], the padded word sequences are transformed into word vectors of 300 dimensions through a word embedding layer.

2. Architecture Details

Audio Encoder E_a. The audio encoder is a ResNetSE34 borrowed from [5]. Specifically, we define the features output from ResNet Stage-2 as shallow feature map, features output from ResNet Stage-3 as middle feature map, features output from ResNet Stage-4 as deep feature map. Then, a series of upsampling, convolution, batchnorm and linear layers transform corresponding audio feature maps into the same size. When the input audio mel-spectrogram of size $1 \times 128 \times 70$, the channel dimension and frequency, time resolutions of different level features $\boldsymbol{f}_a^{\text{low}}$, $\boldsymbol{f}_a^{\text{mid}}$ and $\boldsymbol{f}_a^{\text{high}} \in \mathbb{R}^{32}$ with their corresponding operations are shown in Table 2. The detailed feature dimensions after each operation are shown in Table 3. In this way, the hierarchical audio features are transformed into the same shape and the time dimension is exactly the frame number of a clip, *i.e.*, 34 in our experiment, which is convenient for RNN-based model to take information of each time step as input. After the linear blending of multi-level features, hierarchical audio features for different levels of body parts are established and finally feed to cascaded bi-GRU for pose generation in a coarse-to-fine manner.

Text Encoder E_t. With the speech text pre-processing mentioned above, the word sequences are transformed into word vectors. Next, these word vectors are encoded by an off-the-shelf temporal convolutional text encoder [2]. The text encoder E_t is 4-layered, the receptive field is 16 padded words centered at the current time step and the output dimension of text feature $\boldsymbol{f}_t = E_t(\mathbf{t})$ is 32. In this way, the high-level audio feature and text feature at time-step t are both of dimension 32, which enables the later contrastive learning strategy to leverage the natural audio-text correspondence for achieving discriminative cross-modal feature extraction.

Speaker Identity Encoder E_{ID}. The speaker identity encoder network E_{ID} uses the standard ResNet-18-S5 model. And we do not load ImageNet pretrained weights. We also remove *Global Average Pooling* (GAP) and the final classifier to only leave the visual backbone for visual feature extraction. The input of the encoder is the first M reference frame images of a clip $\{I_1, \ldots, I_M\}$ and the output is speaker identity embedding $\boldsymbol{f}_{id} = E_{\text{ID}}(I_1, \ldots, I_M) \in \mathbb{R}^{18}$. Then through a linear layer of FC (18, 18) and softmax function, $\boldsymbol{f}_{id}$ is transformed into the style coordinator $C \in \mathbb{R}^{3 \times H}$, where $\sum_{i=1}^{3} C[i, h] = 1$. In this way, the speaker identity can affect hierarchical audio feature weight on motion hierarchies.

Motion Hierarchy Establishment. The skeleton of human body is like a tree structure, where the father-joint carries the child-joint to move. To effectively learn the dynamic patterns of different human body parts, we propose to detach the joints from human body ends (fingers) to the main structure (spine) step-by-step and build a motion hierarchy of totally 6 hierarchies: 1) Nose, neck, spine and left/right eye, ear, shoulder; 2) Add left and right elbow; 3) Add left and right wrist; 4) Add left and right finger's first joint; 5) Add left and right finger's second joint; 6) Add left and right finger's third joint. Note that we do not separate more de-

[2]ExPose License: https://github.com/vchoutas/expose/blob/master/LICENSE; OpenPose License: https://github.com/CMU-Perceptual-Computing-Lab/openpose/blob/master/LICENSE; Trimodal: https://github.com/ai4r/Gesture-Generation-from-Trimodal-Context/blob/master/LICENSE.md

Feature Map	Output	Shape	Operation	Feature
Input	-	$1 \times 128 \times 70$	Pre-Conv	-
Shallow	Stage-2	$64 \times 64 \times 35$	Conv2d, ReLU, BN, FC	$\boldsymbol{f}_a^{\text{low}}$
Middle	Stage-3	$128 \times 32 \times 18$	PixelShuffle, Conv2d, ReLU, BN, FC	$\boldsymbol{f}_a^{\text{mid}}$
Deep	Stage-4	$256 \times 16 \times 9$	PixelShuffle, Conv2d, ReLU, BN, FC	$\boldsymbol{f}_a^{\text{high}}$

Table 2. **Definitions of multi-level audio feature maps and transformed features on ResNetSE34.** The shallow/middle/deep feature maps are from the output of ResNetSE34's stage2/3/4. Then, they are transformed by a series of operations to low/mid/high-level audio features, respectively.

Shallow Festure Map from Stage-2	
Operations	Feature Map Shapes
Input	$64 \times 64 \times 35$
Conv2d (64, 2, 1)	$64 \times 63 \times 34$
ReLU, BatchNorm2d (64)	$64 \times 63 \times 34$
Reshape	4032×34
FC (4032, 32)	32×34
Middle Festure Map from Stage-3	
Operations	Feature Map Shapes
Input	$128 \times 32 \times 18$
PixelShuffle (2)	$32 \times 64 \times 36$
Conv2d (32, 3, 1)	$32 \times 62 \times 34$
ReLU, BatchNorm2d (32)	$32 \times 62 \times 34$
Reshape	1984×34
FC (1984, 32)	32×34
Deep Festure Map from Stage-4	
Operations	Feature Map Shapes
Input	$256 \times 16 \times 9$
PixelShuffle (4)	$16 \times 64 \times 36$
Conv2d (16, 3, 1)	$16 \times 62 \times 34$
ReLU, BatchNorm2d (16)	$16 \times 62 \times 34$
Reshape	992×34
FC (992, 32)	32×34

Table 3. **Detailed feature shape after specific operations for the multi-level audio feature extraction.** †Note that in the table, Conv2d (c, k, s) means the output of the convolution is c, kernel size is k and the stride is s; ReLU, BatchNorm2d (c) means the relu and batch-norm operation on the feature of channel size c; Reshape operation combines the channel and frequency dimension together; FC (i, o) means the fully connected linear layer whose input dimension is i and output dimension is o; PixelShuffle (r) means the pixel shuffle operation [16] with resolution r. Specifically, PixelShuffle (r) transforms a feature map of shape $(r^2C) \times H \times W$ into $C \times (rH) \times (rW)$.

tailed motion hierarchy inside the skeleton of human head. Because there is hardly internal movement in the skeleton of head part, *i.e.*, nose, left/right eye and ear. The dynamic among the keypoints of head part is quite trivial under our setting, so it is unreasonable to separate them into different levels of motion hierarchies.

Coarse-to-Fine Pose Generator. According to the established motion hierarchy, the number of keypoints for 6 hierarchies is 9, 11, 13, 23, 33, 43. Since the poses are processed into 3D unit directional vectors representation, the pose dimension of 6 hierarchies is 24, 30, 36, 66, 96, 126. The cascaded hierarchical pose generator contains 6 bi-directional GRU, with the input of corresponding hierarchy pose and audio feature and the hidden size of 300. Note that for the first motion hierarchy, the poses of the first M frames serve as initial poses and are denoted as $\hat{\mathbf{p}}^0 = \{\boldsymbol{p}_1^0, ..., \boldsymbol{p}_M^0, 0, ..., 0\}$; for the later hierarchies, the output from the last pose hierarchy is leveraged to initialize corresponding keypoints. Therefore, $d_s = 300$, $d_a = 32$, $d_p^0 = 24$, $d_p^1 = 24$, $d_p^2 = 30$, $d_p^3 = 36$, $d_p^4 = 66$, $d_p^5 = 96$ and $d_p^6 = 126$ for the parameters W^h and $\boldsymbol{b}^h$ of hierarchical GRU in the Eqn. 4 of main document. In this way, the next pose hierarchy can generate pose with information from the last level of pose, facilitating fine-grained correspondences between audio sequence and co-speech gestures in a coarse-to-fine manner. The last layer's output $\hat{\mathbf{p}}^H$ from the hierarchy is our desired result.

GAN Discriminator D. The architecture of discriminator D is borrowed from [18], with its detailed network design in the Table 4.

Pose Auto-Encoder. Since the generation is a multi-modality problem, it is difficult to use evaluation metrics like L1 distance or L2 distance to judge whether the generated result is good or not. Fréchet Inception Distance (FID) [8] is widely leveraged to evaluate the image generation quality. It firstly pre-train a feature extractor to extract image latent features, then calculates the Fréchet distance between the distributions of the latent feature space of real and generated images. The feature vectors contain more information about characteristics, which is more perceptually plausible than raw pixel space. Based on this, Yoon *et al.* [18] propose a similar evaluation metric Fréchet Gesture Distance (FGD) to evaluate gesture quality.

To further evaluate the pose with expressive finger movements, we train a pose auto-encoder with 43 keypoints on TED-Expressive dataset. The auto-encoder firstly maps 34-frame poses into latent dimension of 128, and then reconstruct them. The detailed structure is borrowed from [18] and recorded in Table 5.

3. Training Stage and Inference Stage

At the training stage, the speech audios, transcripts and reference frames are all needed. The speech audio $\mathbf{a}$ is encoded by the hierarchical audio encoder E_a to get multi-level audio features $\boldsymbol{f}_a^{\text{low}}$, $\boldsymbol{f}_a^{\text{mid}}$ and $\boldsymbol{f}_a^{\text{high}}$. The speech transcript $\mathbf{t}$ is encoded by E_t into text features $\boldsymbol{f}_t$, which are then used by contrastive learning strategy to achieve more discriminative audio feature extraction. Therefore, speech text takes an auxiliary effect in our proposed framework. The reference frames $\mathbf{I} = (I_1, \ldots, I_M)$ are encoded by E_{ID} to represent speaker's identity $\boldsymbol{f}_{id}$, which is then transformed to style coordinator C for feature blending. Besides, reference frames are also used to extract initial poses and finally feed into cascaded bi-GRU to generate co-speech gestures in a coarse-to-fine manner.

At the inference stage, the speech transcript is not needed. **This is the reason why we do not involve the variable $\mathbf{t}$ in the Eq. 1** in our main text. If the reference frames and initial poses are available, we can follow the whole pipeline to generate gestures. For the situations where reference frames and initial poses are unavailable, we can sample a style vector from normal distribution to serve as speaker identity $\boldsymbol{f}_{id}$. Then we can sample an arbitrary sequence of initial poses from the dataset to generate the gestures.

4. Statistics in Physical Constraint

Previous methods on co-speech gesture generation mostly fail to consider human physics constraints, which contributes to unnatural pose and incoherent results. Therefore, we propose to add restrictions on the included angle between bones to ensure reasonable human pose. Concretely, the pose is represented as bone direction vector, which is rendered as $\boldsymbol{p} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_{J-1}]$ and J is the total number of joints. For the j-th bone vector $\boldsymbol{d}_j \in \mathbb{R}^3$ and the (j+1)-th bone vector $\boldsymbol{d}_{j+1} \in \mathbb{R}^3$, we can compute their included angle θ_j by the arc-cosine function on their cosine value. Since there is no benchmark dataset with accurate finger keypoints annotations *under co-speech settings*, we use the hand pose estimator ExPose [4] to annotate the TED-Expressive dataset. With the pseudo ground truth, we can calculate the mean and variance of each angle, which later serve as the mean and variance of Gaussian distribution. The loss function for the physics constraint is the log-likelihood function:

$$\mathcal{L}_{\text{phy}} = -\log \prod_{j=1}^{J-1} \mathcal{N}(\theta_j; \mu_j, \sigma_j^2) = -\sum_{j=1}^{J-1} \log \mathcal{N}(\theta_j; \mu_j, \sigma_j^2), \tag{1}$$

where $\theta_j = \arccos \frac{\boldsymbol{d}_j \cdot \boldsymbol{d}_{j+1}}{\|\boldsymbol{d}_j\| \|\boldsymbol{d}_{j+1}\|}$ is the j-th angle value, μ_j and σ_j^2 are the mean and variance of the j-th angle respectively. We illustrate in Table 6 the means and variances of the included angles (0-180 degrees) around two important joints. In particular, we use θ_s to denote the included angle around the shoulder joint and θ_e for the included angle around the elbow joint. Although some angles may not strictly follow the Gaussian distribution, the intention of physical constraint is to prevent outlier predictions. Thus the assumption of Gaussian distribution could play the role in regularizing generated poses. The ablation study in Table 4 (the setting of "w/o $\mathcal{L}_{\text{phy}}$") shows the effectiveness of such a constraint.

Statistics	Left θ_s	Left θ_e	Right θ_s	Right θ_e
mean(°)	116.6	75.1	127.1	85.3
var(°)	9.01	7.30	7.53	7.22

Table 6. Statistics of important important joint angles.

5. Analysis on Beat Consistency Score Metric

Beat Consistency Score (BC) is a metric adapted by us for motion-audio beat correlation. Previous methods detect motion beats by finding the local optima of kinematic velocity [13], while we propose to utilize the change of included angle between bones to track motion beats. The main reasons are two-fold: 1) previous methods are under the setting of music2dance, where human body involves a global body translation in a large scale. In other word, all of the human's body joints move fast when people dance and the velocity quickly drops when they stop to match a music beat. While in our co-speech gesture settings, the arms are comparatively still and the fingers are more flexible, so their moving scales vary a lot, we can not directly sum up them. 2) Compared to using the shifts of keypoints, we propose to use the included angle to detect motion beat. This is because the human body follow a tree structure. If the arm moves, hand and wrist will follow the movement of arm, which is similar to the process of orbital revolution and self rotation: the father-joint carry the child-joint to move like the orbital revolution and the internal movement of child-joint resembles self rotation. Therefore, directly calculating the Euclidean distance for each joint is unreasonable.

After calculating mean absolute angle change (MAAC) of angle θ_j, we can calculate the sum angle change rate of a certain frame t for the n-th clip as:

$$\frac{1}{J-1} \sum_{j=1}^{J-1} \frac{\|\theta_{j,n,t+1} - \theta_{j,n,t}\|_1}{\text{MAAC}(\theta_j)}. \tag{2}$$

Then we propose to extract the kinematic beat through filtering the angle change rate by following conditions: 1) The angle change rate is a local optimum, *e.g.*, the angle change rate of 9, 10, 11 time-step is 0.2, 0.1, 0.2, respectively. Then the time-step 10 is a local optimum. 2) The difference of

Discriminator D		
Feature	Feature Shapes	Operations
Input	34×126	Transpose (0, 1)
Pre-Conv Layer-1	126×34	Conv1d (126, 16, 3), BatchNorm1d (16), LeakyReLU (0.2)
Pre-Conv Layer-2	16×32	Conv1d (16, 8, 3), BatchNorm1d (8), LeakyReLU (0.2)
Pre-Conv Layer-3	8×30	Conv1d (8, 8, 3), Transpose (0, 1)
Bi-Directional GRU	28×8	Bi-Directional GRU (8, 64)
FC-1	28×64	FC (64, 1), Squeeze(1)
FC-2	28	FC (28, 1), Sigmoid
Output	1	-

Table 4. **Detailed structure and feature shape of Discriminator D.** †Note that in the table, the meanings of contents in operations column are: Conv1d (in_channels, out_channels, kernel_size), BatchNorm1d (feature_dim), LeakyReLU (alpha), Transpose (axis1, axis2), Bi-directional GRU (in_size, hidden_size), FC (in_size, out_size), Squeeze (axis), Sigmoid.

Pose Encoder		
Feature	Feature Shapes	Operations
Input	34×126	Transpose (0, 1)
Layer-1	126×34	Conv1d (32, 3, 1), BatchNorm1d (32), LeakyReLU (0.2)
Layer-2	32×32	Conv1d (64, 3, 1), BatchNorm1d (64), LeakyReLU (0.2)
Layer-3	64×30	Conv1d (64, 4, 2), BatchNorm1d (64), LeakyReLU (0.2)
Layer-4	64×14	Conv1d (32, 3, 1)
Out1	32×12	Flatten, FC (384, 256), BatchNorm1d (256), LeakyReLU (0.2)
Out2	256	FC (256, 128), BatchNorm1d (128), LeakyReLU (0.2), FC (128, 128)
Latent	128	-
Pose Decoder		
Feature	Feature Shapes	Operations
Input	128	FC (128, 64), BatchNorm1d (64), LeakyReLU (0.2), FC (64, 136)
reshape	136	Reshape (4, 34)
Layer-1	4×34	ConvTranspose1d (32, 3, 1), BatchNorm1d (32), LeakyReLU (0.2)
Layer-2	32×36	ConvTranspose1d (32, 3, 1), BatchNorm1d (32), LeakyReLU (0.2)
Layer-3	32×38	Conv1d (32, 3, 1)
Layer-4	32×36	Conv1d (126, 3, 1), Transpose(0, 1)
Pose	34×126	-

Table 5. **Detailed structure and feature shape of Pose Auto-Encoder.** †Note that in the table, Conv1d/ConvTranspose1d (c, k, s) means the output of the convolution/transpose-convolution is c, kernel size is k and the stride is s; LeakyReLU, BatchNorm1d (c) means the leaky-relu and batch-norm operation on the feature of channel size c.

the local optima with either side time-step is larger than a threshold. This is to filter the trivial situation where angle change rates are almost the same during a period of time and guarantee a sudden change of angle change rate as motion beat. For example, the angle change rate of 8, 9, 10, 11, 12 time-step is 0.11, 0.1, 0.11, 0.1, 0.11. It improper to take the time-step 9 and 11 as motion beat. The threshold controls what extent of angle change rate difference is perceived it as a motion beat. A very low threshold will detect the near-stationary motion sequence as many motion beats if there are many trivial beats of type 2 mentioned in the last paragraph. A very high threshold will ignore the normal motion beat. We present the influence of threshold over all baseline method in Fig. 1. We can see that our method can achieve superior performance on BC metric with high robustness to threshold compared to baseline methods. Note that both Attention Seq2Seq [19] and Joint Embedding [1] show low value of BC Score over all threshold, which also proves that they fail to generate results that are synchronous to speech since their gestures are almost still. Although Speech2Gesture [6] shows higher performance on low threshold, they match the trivial beats and perform lower than our method on normal thresholds.

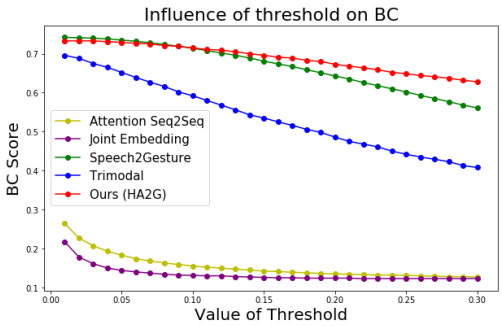

Figure 1. **The influence of threshold on Beat Consistency (BC) Score metric.** We present the BC value of baseline methods and ours under the threshold of range 0.01 to 0.3 with step size of 0.01.

6. Choice of Speaker Identity Extraction

We leverage RGB frames rather than poses for identity information extraction. The appearances of different identities would vary significantly. Though dynamic information can hardly be inferred from the M inputs, our method focuses more on the appearance information like the speaker's height, age and nationality. Inferring speaking styles from appearances or identities only has also been proven effective in Yoon *et al.* [18]. The only speaker-related information we can access is the initial frames, thus we are trying to make the best use of them.

7. Additional Experiments

7.1. Ablation Study on TED Gesture Dataset

We further conduct ablation study on TED Gesture and report results below, which shows the effectiveness of each module. The TED Gesture dataset lacks finger annotations, resulting in the lower motion hierarchy and less significant performance improvement brought by each module.

metric\setting	$\boldsymbol{f}_a^{\text{high}}$ only	Holistic	w/o $\mathcal{L}_{\text{phy}}$	HA2G-ASR	**HA2G Full**
FGD	3.569	3.682	3.165	3.091	**3.072**

7.2. Influence of Reference Frame Number M

All the models are implemented with the same amount of information given as the input, including the number of initial poses M. The setting of using $M = 4$ frames as seed pose is proposed in Trimodal [18]. Our whole setting basically follows theirs. To investigate the influence of M, we further set M as 1 and 7. The results below suggest that the performance gain derived from additional initial poses is marginal, which shows the robustness of the proposed method to hyper-parameter M.

M	FGD ↓	BC ↑	Diversity ↑
1	5.994	0.708	169.425
4	5.306	0.715	173.899
7	5.177	0.715	174.313

Table 7. Influence of reference frame number M.

7.3. Randomness of Diversity Metric

The Diversity metric is adapted from [11] and is popularly used in other works [9]. In order to mitigate the influence of randomness, we randomly sample 500 pairs, which is much more than the number 200 in [11]. To ensure and verify the reproducibility, we further conduct the evaluation 10 times (create random samples 10 times with different random seeds). The results are listed in the table below. We can see that the difference is comparatively small between each group, which proves that the Diversity metric can be reproduced and the sample number of 500 is enough to alleviate randomness.

Group	1	2	3	4	5
Diversity	172.58	171.91	173.60	173.66	173.71
Group	6	7	8	9	10
Diversity	172.12	171.88	173.02	173.80	172.83

Table 8. Randomness of the Diversity metric.

8. Limitations and Future Work

Our work mainly have the following limitations: **1)** Since when people talk to others, the most important non-verbal behavior is upper body movements. Hence we only delve into the co-speech gesture generation of human upper body, without considering full body motions. This will make our trained avatars fail to walk around like TED Talk narrators. **2)** In the TED Talk dataset, some data samples have very strong prior on human hand pose at the specific settings that will affect training, *e.g.*, people with speaker or chalk in their hand as shown in Fig. 2. **3)** Although our proposed approach can capture the fine-grained motions of co-speech finger movements and diverse dynamic patterns of different human body parts, we still find it difficult to capture some very subtle movements like "shrug". This is mainly due to the fact that there hardly exists such action samples in the dataset and it is very hard for our model to learn such dynamic patterns. In future work, we will improve our method to capture full-body co-speech gestures and some very minor pose movements and we will enhance the automatic dataset pipeline algorithm to filter samples with strong prior that may affect our training quality.

Figure 2. **Examples of data samples at specific setting with very strong prior on hand pose.** We implement the mosaic operations for all the images to eliminate personally identifiable information.

9. Social Impact

Making co-speech gestures to complement conversational information is a kind of innate non-verbal behavior for human, while this work encourages the machine intelligence to be equipped with such ability, especially learn to animate the subtle hand and arm motions. Therefore, this work can exert positive impacts on both machine learning research and application field. On the one hand, the proposed approach identifies the advantages of hierarchical architecture design to extract cross-modal information at multiple granularities and excavate the fine-grained audio-pose associations, which can further facilitate cross-modal animation tasks like talking face generation and music2dance prospectively. On the other hand, the speech-driven gesture generation technique has a wide range of beneficial applications for society, including digital human broadcaster and social robots. Specifically, it could also assist dumb people to learn communication skills by teaching sign language with expressive human-like motions. Since the generated motions are all skeleton-based, they hardly have detrimental impact in most cases. Still, it may potentially lead to the misuse of copyrighted 3D character models if we animate them without permission. Besides, the bias of the dataset may have some negative impact, *e.g.*, some gestures may have negative meanings for some nations. But we believe the proper use of this technique will enhance positive societal development.

10. Details of User Study

The study involves 24 participants. They take 25-35 min to complete the task. The participants are 12 females and 12 males, with age range of 18-24 years old. The users are unaware of which motion sequence corresponds to which method or even the ground truth. Specifically, we randomly shuffle the order of video placement for all methods every time, so that participants can concentrate only on the quality of generated results for fair comparison.

We have provided the users with instructions before conducting the study. The participants are asked to judge the three perspectives in the following manner: a) For "Naturalness", does motion look natural and like real human poses regardless of background speech? There should not be any strange angles and unnatural movements. b) For "Smoothness", does the generated motion maintain smoothness in temporal dimension, with no obvious rigid or stuck movements, regardless of background speech audio? c) For "Synchrony", does the generated motion match the corresponding speech audio both rhythmically and semantically? We also show the raw videos of TED Talk before participants' rating process to help them make more accurate judgement. 24 participants of 12 females and 12 males are involved in the study, covering 4 nationalities in order to bridge biases. 12 of them are researchers from the field of deep generative models and others are from other fields.

References

[1] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE, 2019. 5

[2] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018. 2

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019. 1, 2

[4] 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)*, 2020. 1, 2, 4, 8

[5] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han. In defence of metric learning for speaker recognition. *arXiv preprint arXiv:2003.11982*, 2020. 2

[6] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019. 1, 2, 5

[7] 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. *arXiv preprint arXiv:2102.06837*, 2021. 1

[8] 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. In *Neural Information Processing Systems (NIPS)*, 2017. 3

[9] Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. Dance revolution: Long-term dance generation with music via curriculum learning. *arXiv preprint arXiv:2006.06119*, 2020. 6

[10] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S Srinivasa, and Yaser Sheikh. Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 763–772, 2019. 1

[11] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 6

[12] Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. Gesture controllers. In *ACM SIGGRAPH 2010 papers*, pages 1–11. 2010. 1

[13] Buyu Li, Yongchi Zhao, and Lu Sheng. Dancenet3d: Music based dance generation with parametric motion transformer. *arXiv preprint arXiv:2103.10206*, 2021. 4

[14] Ochshorn Robert and Hawkin Max. Gentle: A forced aligner. 2016. 2

[15] Najmeh Sadoughi, Yang Liu, and Carlos Busso. Msp-avatar corpus: Motion capture recordings to study the role of discourse functions in the design of intelligent virtual agents. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 7, pages 1–6. IEEE, 2015. 1

[16] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[17] Jackson Tolins, Kris Liu, Yingying Wang, Jean E Fox Tree, Marilyn Walker, and Michael Neff. A multimodal motion-captured corpus of matched and mismatched extravert-introvert conversational pairs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3469–3476, 2016. 1

[18] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020. 1, 2, 3, 6

[19] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4303–4309. IEEE, 2019. 1, 2, 5

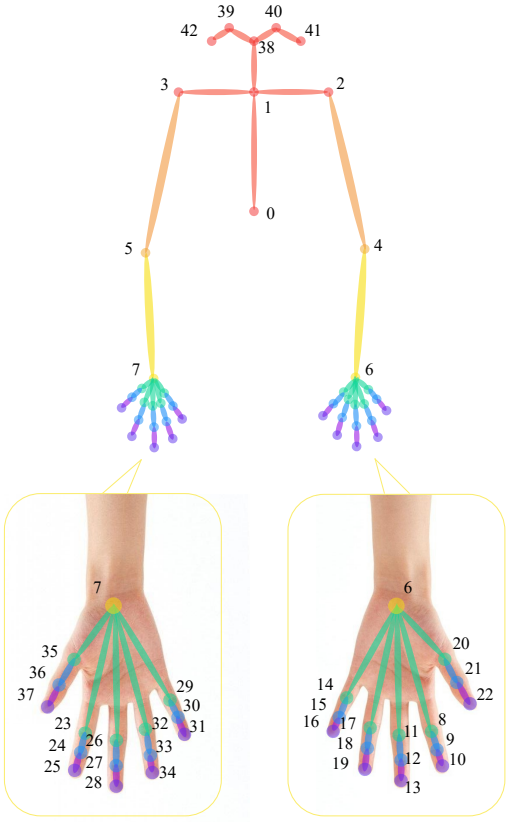

Figure 3. **The detailed 3D keypoints output annotation by ExPose [4].** In particular, we annotate 43 upper body keypoints, including: spine (0), neck (1), left shoulder (2), right shoulder (3), left elbow (4), right elbow (5), left wrist (6), right wrist (7), left index (8, 9, 10), left middle (11, 12, 13), left pinky (14, 15, 16), left ring (17, 18, 19), left thumb (20, 21, 22), right index (23, 24, 25), right middle (26, 27, 28), right pinky (29, 30, 31), right ring (32, 33, 34), right thumb (35, 36, 37), nose (38), right eye (39), left eye (40), right ear (41), left ear (42). Note that the holistic upper body with keypoints index is shown at the top of figure, the zoom-in images of left hand and right hand with detailed annotations are shown at the bottom.