Cultural Self-Adaptive Multimodal Gesture Generation Based on Multiple Culture Gesture Dataset

Jingyu Wu
wujingyu@zju.edu.cn
College of Computer Science and Technology
Hangzhou, Zhejiang, China

Shi Chen*
Zhejiang-Singapore Innovation and AI Joint Research Lab
Hangzhou, Zhejiang, China
shelleych@zju.edu.cn

Shuyu Gan
College of Computer Science and Technology
Hangzhou, Zhejiang, China
derek_sygan@outlook.com

Weijun Li
College of Computer Science and Technology
Hangzhou, Zhejiang, China
mvs@zju.edu.com

Changyuan Yang
College of Computer Science and Technology
Hangzhou, Zhejiang, China
11821114@zju.edu.cn

Lingyun Sun
Zhejiang-Singapore Innovation and AI Joint Research Lab
Hangzhou, Zhejiang, China
sunly@zju.edu.cn

ABSTRACT

Co-speech gesture generation is essential for multimodal chatbots and agents. Previous research extensively studies the relationship between text, audio, and gesture. Meanwhile, to enhance cross-culture communication, culture-specific gestures are crucial for chatbots to learn cultural differences and incorporate cultural cues. However, culture-specific gesture generation faces two challenges: lack of large-scale, high-quality gesture datasets that include diverse cultural groups, and lack of generalization across different cultures. Therefore, in this paper, we first introduce a Multiple Culture Gesture Dataset (MCGD), the largest freely available gesture dataset to date. It consists of ten different cultures, over 200 speakers, and 10,000 segmented sequences. We further propose a Cultural Self-adaptive Gesture Generation Network (CSGN) that takes multimodal relationships into consideration while generating gestures using a cascade architecture and learnable dynamic weight. The CSGN adaptively generates gestures with different cultural characteristics without the need to retrain a new network. It extracts cultural features from the multimodal inputs or a cultural style embedding space with a designated culture. We broadly evaluate our method across four large-scale benchmark datasets. Empirical results show that our method achieves multiple cultural gesture generation and improves comprehensiveness of multimodal inputs. Our method improves the state-of-the-art average FGD from 53.7 to 48.0 and culture deception rate (CDR) from 33.63% to 39.87%.

CCS CONCEPTS

• **Computing methodologies** → **Animation**; **Artificial intelligence**; *Supervised learning*.

*Corresponding author

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611705

KEYWORDS

co-speech gesture generation, datasets, multimodal chatbots, evaluation metric, nonverbal behavior

ACM Reference Format:
Jingyu Wu, Shi Chen, Shuyu Gan, Weijun Li, Changyuan Yang, and Lingyun Sun. 2023. Cultural Self-Adaptive Multimodal Gesture Generation Based on Multiple Culture Gesture Dataset. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3581783.3611705

1 INTRODUCTION

Multimodal co-speech gesture generation is essential for developing social chatbots and agents [21, 42]. The gestures not only complement speech and add non-verbal information that helps listeners to concentrate and understand utterances [10, 13], but also improve the intimacy between humans and multimodal chatbots [65]. However, co-speech gesture generation is a challenging problem because learning and generating gestures have to comprehend the complex relationship between speech, gestures, and other different modalities [42, 73].

Gesture is unique [57], as different cultures [19] influence speakers to use different styles of motion [37]. For example, when referring to the past, English speakers often move their hands backward to indicate the past [6]. Spanish speakers, on the other hand, prefer to swing their arm from the chest to the left [49]. Meanwhile, Chinese speakers tend to swing their arm to the left at a lower level [9]. While recent research on co-speech gesture generation extensively studies the integration and relationship of multiple modalities, such as audio [21, 41], text [4, 5, 8], emotion [8, 42], speaker identity [42, 73] and so on, culture-specific gesture genenration remains unsolved and presents several challenges.

The first challenge comes with the quality and scale of available datasets. Previous methods [46, 60, 67] are mainly trained on English-language multimodal datasets, such as GSD [58], YouTube Gesture [21], and TED Gesture [73]. However, most available datasets, as shown in Table. 1, have either only one speaker [17] or many speakers [21, 23, 62, 73] from same culture (mostly in English). Currently, BEAT [42] is proposed, including four different languages,

Table 1: Comparison of datasets. We compare with several gesture datasets that are used more frequently in the community. "#" indicates the number. The best scores are reported in bold. Our dataset is the largest freely available gesture dataset with multiple cultures and will provide a great contribution to the efforts in developing naturalistic chatbots in different cultures.

Dataset	#Modality	# Speaker	Audio (Culture)	Text	# Sequence	Duration [hrs]
MPI [63]	3	1	-	✓	1,408	1.5
Takechi [58]	3	2	Jp	-	1,049	5.0
YouTube Gesture [21, 23]	5	6	En	-	N/A	33
Taking16.2M [40]	5	50	En	✓	N/A	50
TED Gesture [73, 74]	4	>100	En	✓	1,400	97
BEAT [42]	7	30	En/Cn/Es/Jp	✓	2,508	76
MCGD (Ours)	5	263	En/Cn/Es/Jp/Kr Fr/It/Ru/De/Ar	✓	10,414	103

but over 81% of the data is in English. However, people from different cultures exhibit distinct gesture characteristics in various circumstances, such as agreement [19], rejection [24], and online-talking [70]. Therefore, it is crucial for multimodal chatbots to learn culture-specific gestures when communicating with users from different native cultures. Building a multiple native culture dataset will greatly improve the performance of naturalistic multimodal chatbots facing cross-culture situations.

The second challenge is the lack of capability for generalization and flexibility across different languages. Current approaches are typically designed to learn co-speech gesture characteristics from a collection of videos of a single culture [21, 42, 46, 73]. Facing each new culture, they are required to retrain the model, which can be time- and resource-consuming [39]. Hence, methods that can self-adapt to multiple cultures and generate culture-specific gestures are significant for developing multimodal chatbots.

Motivated by the challenges outlined above, we first introduce a pseudo-label dataset named Multiple Culture Gesture Dataset (MCGD), which contains ten different cultures represented by over 200 speakers across five modalities. Building on the methodology of Lotfian and Busso [44], our data collection and annotation process is flexible and scalable, resulting in 103 hours of audiovisual data and more than 10,000 segmented sequences (Sec. 3). We carefully designed the datasets to cover a wide range of natural language characteristics, including the ratio of male/female, range of phonemes, and variety of languages. Furthermore, we further observed the correlation of gestures with different cultures after statistical analyses on MCGD. Overall, the MCGD is the largest freely available gesture dataset to date, consisting of ten different cultures and over 200 speakers.

Additionally, we propose a Cultural Self-adaptive Gesture Generation Network (CSGN). It learns to synthesize gestures based on the three modalities mentioned above (text, audio, and culture) and adapts to multiple cultures by parameter sharing. The proposed method consists of cascaded encoders and dynamic weights that enhance the contribution of audio and text features. It extracts cultural features from the multimodal inputs of different cultures or a cultural style embedding space with a designated culture. Moreover, we use a classification network to classify latent codes of culture to avoid trivial solutions and encourage each cultural feature to have spatial diversity distance from each other (Sec. 5.3).

Besides, to evaluate the performance of CSGN in generating different cultural gestures, we further propose a Culture Deception Rate (CDR), inspired by the deception rate used in artist-style transfer [14, 56]. The CDR calculates the rate at which the network classifies the generated gestures into the correct cultural features.

Through qualitative and quantitative experiments, our method achieves adaptive culture-specific gesture generation. The generated gestures are unique to different cultures and preserve diversity (Fig. 6 (c)), and increase the state-of-the-art Culture Deception Rate (CDR) by 18.6% (Sec. 5.3). We further compare several models using different generated architecture, such as adversarial training [73], quantization [46], and flow-based [3, 26]. The results show that our method gets state-of-the-art performance and improves the average FGD [73] from 53.7 to 48.0 (Sec. 5.3).

Extensive experiments demonstrate that our approach is highly effective, achieving state-of-the-art performance on four large-scale benchmark datasets. In summary, our main contributions are:

- We release a Multiple Culture Gesture Dataset (MCGD), which is the largest freely available co-speech gesture dataset to date, consisting of ten different cultures, over 200 speakers, 10,000 segmented sequences, and a wide range of natural language characteristics. The MCGD may provide a great contribution to the efforts in developing naturalistic multimodal chatbots in different cultures.
- We propose a Cultural Self-adaptive Gesture Generation Network (CSGN). It adapts to multiple cultures and learns the relationship with audio, text, and culture through cascaded architecture. We further introduce a new indicator Culture Deception Rate (CDR) to evaluate the performance of culture-specific gesture generation.
- We broadly evaluate our approach across four large-scale datasets with several state-of-the-art gesture generation methods. Quantitative and qualitative evaluation results demonstrate the improvements in our method are concise and effective.

2 RELATED WORK

Co-speech Gesture Datasets. The co-speech gesture datasets support the development of multimodal chatbots and agents. These datasets are used to study co-speech gestures in natural settings using data-driven approaches [21, 42, 46, 73]. Additionally, psychologists use these datasets to study various aspects of co-speech gesture [34, 57] (see Wagner et al. [64] for a review).

Existing gesture datasets are annotated from two perspectives: motion capture and pseudo-labeling. The former involves precisely and accurately capturing the motion process through the use of motion capture devices, such as the MPI [63], Takechi [58], and Taking16.2M [40]. Recently, Liu et al. [42] build a large body expression audio-text dataset (BEAT), which captures the co-speech gestures of 30 speakers expressing eight different emotions. The other tries to annotate the gestures with the help of deep learning algorithms. Ginosar et al. [21] create a YouTube Dataset which uses OpenPose [12] to extract 2D key poses from YouTube videos, and Habibie et al. [23] extent it to a full 3D body with facial landmarks. Recently, Yoon et al. [73] build a TED dataset using VideoPose3D [51].

However, the resources currently developed for other cultures are few and of different domains [16, 60, 68]. In fact, between the 359 multimodal resources certified for all languages by the LRE map [52], most languages have only one or two freely available datasets while English has over 100. In order to ensure that multimodal chatbots are able to adapt to local users, it is important to train them on native datasets rather than simply translating text from English datasets [33]. Though BEAT [42] supports four different languages, it only includes data from ten individuals, and over 81% of the data is based on English.

To address this limitation and to increase the dataset's native adaptability and scale, we introduce the Multiple Culture Gesture Dataset (MCGD). This dataset includes data from ten cultures, with each culture having over 20 people to reduce individual bias. We believe the MCGD will make a significant contribution to efforts aimed at developing naturalistic chatbots in different cultures.

Conditional Conversational Gesture Generation. Previous works are released with one or two modalities as conditions, such as text-gesture generation [74], audio-gesture generation [17, 21, 58] and audio-text-gesture generation [42, 73]. Early models are mainly based on CNN [39] or LSTM [28] for end-to-end training. Recently, several efforts try to improve the performance by using generative adversarial networks (GAN) [18, 66, 67], quantization [46], flow-based [3, 26] and various types of synthesis techniques [1, 45].

However, previous research mainly focuses on the text or audio as conditions, ignoring the cultural influence. Recent works on multimodal chatbots and gestural analyses highlight the significant impact of culture on the use of gestures. Trotta and Guarasci [60] analyze the co-gesture behaviors of several Italian and English politicians during face-to-face interviews, while Gander et al. [19] explore how communicative gestures are used to express agreement in first encounters by Swedish and Chinese. Additionally, the cultural influence on other languages is also explored according to recent research, such as Finnish [59] and Danish [50].

Although the Mix-StAGE [3] and MultiContext [73] consider individual speaker's style and video's ID, respectively, the cultural characteristics can differ significantly between individuals. We propose the CSGN which is designed to adapt multiple languages without requiring retraining a new network and achieves state-of-the-art performance across four large-scale benchmark datasets.

3 MCGD: MULTIPLE CULTURE GESTURE DATASET

In this section, we provide a brief overview of the proposed Multiple Culture Gesture Dataset (MCGD). We first describe the data acquisition process and the annotation procedure. Then, we analyze the differences and similarities between different cultures using MCGD, and present some key distributions. Supplementary materials provide more detailed information about our dataset.

3.1 Data Acquisition

The data acquisition process of MCGD is designed to be flexible and scalable to create favorable conditions for generating gestures.

The first step is to select candidate self-speech recordings from multiple video-sharing websites. As suggested by Vidal et al. [62], the videos that are selected need to have frontal body when speaking, clear speech, and no background music. Videos that are outdated, of low resolution, or featuring music performances or interviews are excluded. The second step involves strictly controlling the proportion of languages as well as accents to ensure the generalization capability to different cultures. The videos are then segmented into separate short video groups. The final step in this process is to manually annotate the identities of the participants in the videos. Each participant is given a unique ID number.

In total, we collect 10,414 segments, consisting of ten different cultures. Each culture contains at least 20 native speakers (around half of whom are female) and over ten hours of videos. More distributions of our dataset are shown in Fig. 1, and additional information can be found in the supplementary materials.

3.2 Data Annotation

The annotation of MCDG refers to BEAT [42] and TED Gesture [74]. Multimodal signals are divided into text, audio, culture, and gesture.

For the text translation and alignment, we use an in-house-built Automatic Speech Recognizer (ASR) to translate the initial text from the videos and proofread it by at least two native annotators. The inter-coder reliability checking is performed independently by other annotators. The inter-coder agreement for translation reached a Cohen's *kappa score* in the range 0.75-0.80 [48]. Finally, we adopt Montreal Forced Aligner (MFA) [47] for temporal alignment of the text with audio.

For the gesture annotation, we represent the speakers' pose over time using a temporal stack of 2D skeletal keypoints [21], which we obtain using OpenPose [12]. From the complete set of keypoints detected, we select ten points corresponding to the head, neck, R/L shoulders, R/L elbows, R/L wrists, and R/L hands to represent gestures. We further converted all human poses to 3D by using the 3D pose estimator [51] which converts a sequence of 2D poses into 3D poses. Together with the video footage, we provide the skeletal keypoints for each frame of the data at 15 FPS. Note that these are not ground truth annotations, but rather a proxy for the ground truth obtained from a state-of-the-art pose detection system.

In addition, we replicate the online approach proposed in Lotfian and Busso [44], which is a modified version of the crowd-sourcing protocol presented by Burmania et al. [11]. This approach tracks the performance of workers in real-time, stopping the annotation when their quality drops below a certain threshold.

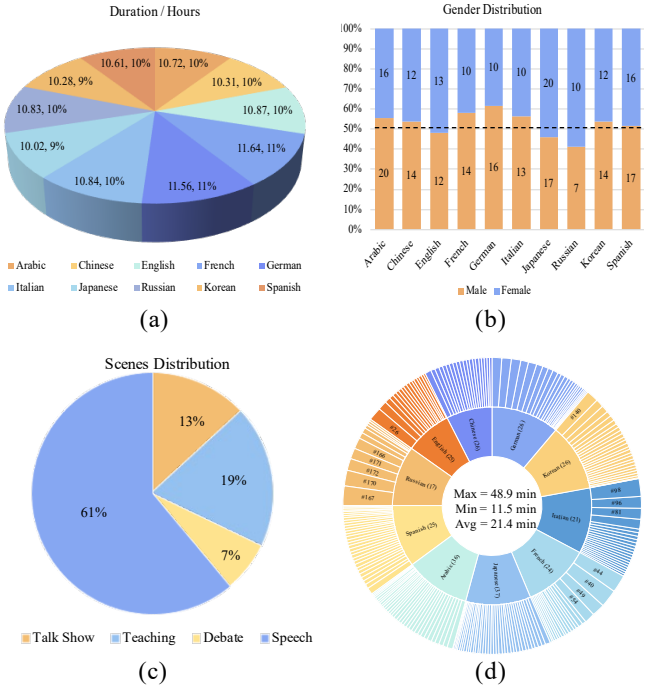

Figure 1: Distributions of our MCGD. (a) Our dataset includes ten cultures, and the proportion of each culture is approximately the same. (b) Gender distribution of each culture. (c) Gestures are divided into four scenes which mainly consist of speech. (d) By 263 speakers from ten cultures with different recording duration.

4 METHOD

What is culture-specific gesture? Specifically, in a given culture, individuals often have their own unique gesture customs, but overall, these gestures share common characteristics. Given raw audio or text of a speech, our goal is to not only generate the speaker's corresponding gesture motion, but also adapt to multiple cultural characteristics without training a new network.

To integrate all the modalities and incorporate cultural features, we propose a cultural self-adaptive gesture generation network (CSGN). The structure of our framework (Fig. 2) contains three main components: (1) a multi-stage, cascade architecture that encodes the input modalities and assigns weights for different modalities dynamically; (2) a cultural self-adaptive module that extracts the characteristics of multiple cultures; (3) an autoregressive transformer decoder for predicting gestures.

4.1 Text and Audio Encoder

Text Encoder. The input text is a sequence of words whose length varies according to the speed of speech. We first insert padding tokens ($\diamond$) into the word sequence to make it the same length as gestures [73]. Next, the input text is passed by a culture identification model [31, 32] and then the word in each frame is converted to word embedding $\mathbf{w}^T \in \mathbb{R}^{300}$ by the pre-trained model in FastText-multilingual [22] to reduce dimensions. After that, the customized text encoder E_T, an 4-layer temporal convolution network (TCN) [7] with skip connections [25], extracts the features of the text $\mathbf{Z}^T$:

$$\mathbf{W}^T = (\mathbf{w}^T_{i-f}, ..., \mathbf{w}^T_{i+f}) \tag{1}$$

$$\mathbf{Z}^T = (\mathbf{z}^T_{i-f}, ..., \mathbf{z}^T_{i+f}) = E_T(\mathbf{w}^T_{i-f}, ..., \mathbf{w}^T_{i+f}) \tag{2}$$

For each frame i, the encoder E_T generates the final latent features of text by fusing information from $2f = 34$ frames. The set of word embedding and features are denoted as $\mathbf{W}^T \in \mathbb{R}^{2f\times 300}$ and $\mathbf{Z}^T \in \mathbb{R}^{2f\times 128}$, respectively.

Audio Encoder. Since the length of audio input is usually fixed, we represent audio using its raw waveform. We downsample the audio to 16K HZ, considering the audio as 15 FPS. The audio encoder E_A extracts the audio feature $\mathbf{Z}^A \in \mathbb{R}^{2f\times 128}$ as:

$$\mathbf{Z}^A = (\mathbf{z}^A_{i-f}, ..., \mathbf{z}^A_{i+f}) = E_A(audio, Mel(audio)) \tag{3}$$

Where the $Mel()$ represents the Mel Spectrogram of audio.

Additionally, to assign weights to audio and text and to enhance their relevance, we fuse the $\mathbf{Z}^A$ and $\mathbf{Z}^T$ using a learnable attention mechanism. Specifically, we compute the attention weight γ_i for each frame i as:

$$\gamma_i = \frac{exp(\mathbf{W}_a\mathbf{z}^A_i + \mathbf{W}_t\mathbf{z}^T_i + b)}{\sum_{j=-f}^{f} exp(\mathbf{W}_a\mathbf{z}^A_{i+j} + \mathbf{W}_t\mathbf{z}^T_{i+j} + b)} \tag{4}$$

where $\mathbf{W}_a$ and $\mathbf{W}_t$ are learnable weights and b is a bias term. We compute the fused feature $\mathbf{z}^{AT}_i$ as the weighted sum of $\mathbf{z}^A_i$ and $\mathbf{z}^T_i$:

$$\mathbf{z}^{AT}_i = \gamma_i\mathbf{z}^A_i + (1-\gamma_i)\mathbf{z}^T_i \tag{5}$$

the resulting feature $\mathbf{Z}^{AT} = (\mathbf{z}^{AT}_{i-f}, ..., \mathbf{z}^{AT}_{i+f})$ has the same shape as $\mathbf{Z}^A$ and $\mathbf{Z}^T$, and contains information from both modalities

4.2 Cultural Self-adaptive Module

Our cultural self-adaptive module has two parts: (1) the culture characteristic extractor E_C; (2) the cultural embedding space.

The extractor E_C extracts the cultural feature $\mathbf{Z}^C \in \mathbb{R}^{2f\times 64}$ from the word embedding set $\mathbf{W}^T$ and the penultimate layer's output $\mathbf{Z}^{A'}$ of audio encoder E_A.

$$\mathbf{Z}^C = P(\mathbf{z}^C) = P(E_C(\mathbf{W}^T, \mathbf{Z}^{A'})) \tag{6}$$

To ensure the correct decomposition of cultural feature $\mathbf{Z}^C$ and avoid trivial solutions, a classification network CCN_1 is used to classify the culture. We add an average pooling layer before classification to prevent E_C from homogenizing the culture features and to preserve diversity [14]. Experiments (Fig. 6) demonstrate our method is capable of achieving culture-specific gesture generation.

Furthermore, facing users who want to replace cultural features from input modalities with another culture, we add a cultural style embedding space, which converts cultural style to another or a specific style at the synthesis phase. It utilizes the culture IDs to reflect the characteristics of each culture in the dataset. The cultural style embedding space is a pre-trained and comprehensive feature space, which is learned from our dataset. Each culture in the embedding space is not unique. Meanwhile, to make the cultural features in embedding space more interpretable, variational inference [36, 54] that uses a probabilistic sampling process is used. The training and structure details of cultural embedding space are in the appendix.

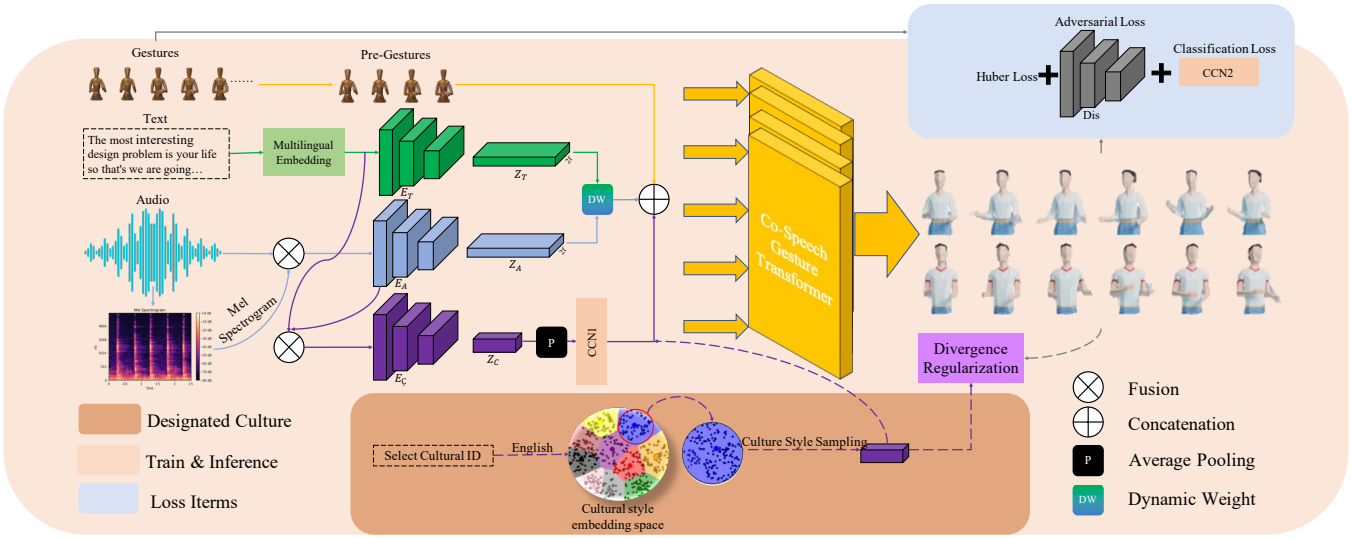

Figure 2: The overview of the proposed co-speech gesture generation network that generates the speaker's corresponding gesture motion and adapts to multiple cultural characteristics without training a new network. With a cascaded architecture and learnable dynamic weights, it enhances the relationship between text and audio. Furthermore, it extracts cultural features from the multimodal inputs or a cultural style embedding space with a designated culture.

4.3 Gesture Decoder

The gesture decoder G takes encoded features of audio and text as input, after fusing the cultural characteristic, and then generates gestures. Instead of taking a sequence of human poses as gestures, we represent each pose as directional vectors which represent the relative positions of the child joints from the parent joints and are favored for training [73]. We have nine directional vectors (e.g., spine-neck, neck-R/L shoulders, etc.) and each vector has three values indicating the 3-dimensions.

Specifically, for gesture generation, we use a transformer architecture which is a customized conditioned GPT-2 [71] with mask self-attention. For each training sample, the decoder G predicts a sequence of gestures according to the encoded features of text, audio, culture characteristics, and pre-gestures:

$$\mathbf{Z}^M = (\mathbf{z}^M_{i-f}, ..., \mathbf{z}^M_{i+f}) = Fusion(\mathbf{Z}^{AT}, \mathbf{Z}^C) \tag{7}$$

$$\hat{\mathbf{V}} = (\hat{\mathbf{v}}_{i-f}, ..., \hat{\mathbf{v}}_{i+f}) = G(\mathbf{Z}^M, gestures_{pre}) \tag{8}$$

Where the final estimated gesture $\hat{\mathbf{V}} \in \mathbb{R}^{2f \times 27}$ and $\mathbf{Z}^M \in \mathbb{R}^{2f \times 300}$ is the fused features for all modalities. For Eq. 8, the length for pre-gestures is the initial four frames.

4.4 Overall Training

For the performance of gesture generation, we use Huber loss [29]:

$$\mathcal{L}_{rec} = \mathbb{E}[\frac{1}{2f} HuberLoss(\mathbf{V}, \hat{\mathbf{V}})] \tag{9}$$

However, Huber loss suffers from the known issue of regression to the mean which produces overly smooth motion [21]. Therefore, we further add an adversarial loss to combat it and ensure the realistic of gestures:

$$\mathcal{L}_{GAN,adv} = -\mathbb{E}[\log(D(G(\mathbf{Z}^M, gestures_{pre})))] \tag{10}$$

where the discriminator D input to the adversarial training is only the gesture itself. Additionally, inspired by Wu et al. [69], we add a culture classification network (CCN_2) to encourage generator G to produce gestures that are culturally diverse and can be distinguished from each other. The loss function of CCN_1 and CCN_2:

$$\mathcal{L}_{cls} = \alpha \mathcal{L}_{ce}(CCN_1(P(\mathbf{z}^C)), l) + (1-\alpha)\mathcal{L}_{ce}(CCN_2(\hat{\mathbf{V}}), l) \tag{11}$$

where P, l and $\mathcal{L}_{ce}$ represent the average pooling layer, label and cross-entropy loss, respectively. We also adopt a weight α to balance the weight of CCN_1 and CCN_2. After that, during training, we adjust the weights of each loss and the final loss function is:

$$\mathcal{L} = \beta_0 \mathcal{L}_{rec} + \beta_1 \mathcal{L}_{GAN,adv} + \beta_2 \mathcal{L}_{cls} \tag{12}$$

where the weights for loss terms β_0, β_1 and β_2 are determined experimentally. The detailed implementations of our networks are in the supplementary materials.

4.5 Culture Deception Rate (CDR)

While several metrics [21, 42, 73] are proposed to evaluate the quality of gesture generation, until now no evaluation metric has been proposed for an automatic evaluation of specific-style gesture generation. We propose the culture deception rate (CDR) to evaluate the cultural relevancy of gestures, which is inspired by deception rate [56] from style transfer [20]. To compute the CDR, we use a customized ResNet [25] for time series classification as backbone and trained from scratch to classify gestures from different cultures using our MCGD. The culture deception rate is calculated as the fraction of generated gestures that are correctly classified by the network. To ensure a fair comparison, we train the other methods for different cultures separately. The experiments for CDR are conducted on Sec. 5.3. More details are shown in the appendix.

5 EXPERIMENTS

We conduct extensive experiments to evaluate our method, including qualitative comparisons of gesture generation results synthesized by our model and other baselines in Section 5.2, followed by presenting quantitative results in Section 5.3. Finally, in Section 5.4, we conduct ablation studies to validate the effectiveness of each component.

5.1 Experimental Setup

Datasets We evaluate the performance on the following three large datasets for single-cultural gesture generation: (1) YouTube Gesture [21], (2) TED Gesture [74], and (3) BEAT [42]. For multiple cultural gesture generation, we use MCGD for training and testing. All gesture poses are resampled at 15 frames per second and each training sample has 34 frames which are sampled with a stride of 10 from the video sections. Note that in the inference stage, the initial four frames are used as seed poses and the models are trained to generate the remaining 30 poses (2 seconds).

Baselines To evaluate the performance of our method, we compare several state-of-the-art architectures, such as Seq2Seq [74], Speech2Gesture [21], Joint-Embedding [2], Audio2Gesture [41], MultiContext [73],CaMN [42] and ATG [46]. Among them, Seq2Seq [74], Joint-Embedding [2] and Audio2Gesture [41] are based on the CNN or LSTM for end-to-end training. Speech2Gesture [21] uses adversarial training to improve performance. MultiContext [73] and CaMN [42] learn features with cascaded network architectures. And ATG [46] combines the vector quantisation (VQ-VAE) [61] with GPT [71] as a two-stage method. All baselines are trained using publicly available implementations with default configurations.

5.2 Qualitative Results

We qualitatively compare our co-speech gesture generation results to the baselines in Fig. 3. More results of our method under different conditions are shown in Fig. 4 and Fig. 5. The qualitative results in Fig. 4 show some common characteristics of gestures. For instance, gesture generation depends on speech rhythm and presence or absence of speech in (a). Additionally, when referring to a range (e.g., from A to B, between A and B, etc.), gestures tend to swing from one direction to the other, such as left to right in (b). Furthermore, when encountering strong emotional expressions, the model often synthesis the gestures that open the arms as shown in (c).

While retaining common characteristics of gestures, we achieve differences across different cultures. We find that compared with other cultures, the gestures generated by Arabic prefer to use the right hand for more motions with the same multimodal inputs as shown in Fig. 5 (a) (b). Because the concept of the left hand is considered unclean in Arab culture [15, 38]. While speaking of "I", all cultures show relatively high consistency, typically using the left, right, or both hands to point towards oneself, as shown in Fig. 5 (c). When talking about the future, generated by English prefer to move their hand to the left, by Chinese prefer to swing arm to the left, and Arab pinch virtual timelines with their fingers and slide to the left in Fig. 5 (d).

All gestures are depicted using a 3D dummy character. The poses represented as directional vectors are retargeted to the character with fixed bone lengths, and the gesture sequences are upsampled using cubic spline interpolation to 30 FPS. We used the same retargeting procedure for all animations. Please refer to our supplementary video results which better convey temporal information.

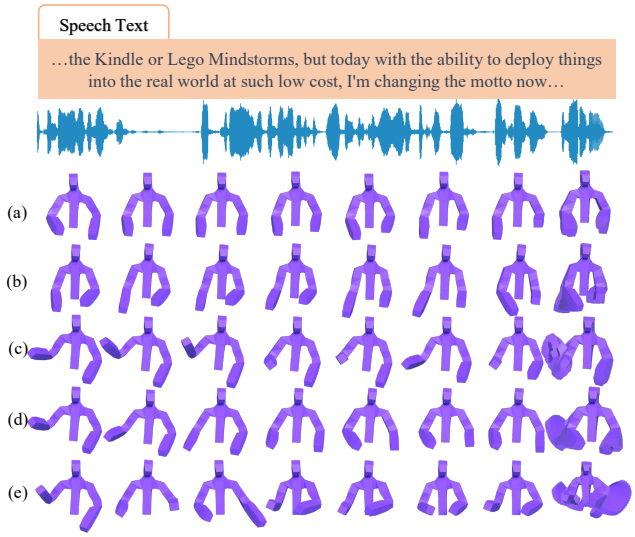

Figure 3: Qualitative comparisons of (a) Joint-Embedding, (b) MultiContext, (c) ATG, (d) CaMN, and (e) Ours for the same input speech. The first seven images show seven evenly sampled frames from the resulting pose sequences. The last column shows motion history images in which all frames are superimposed. Please see the supplementary video for animated results.

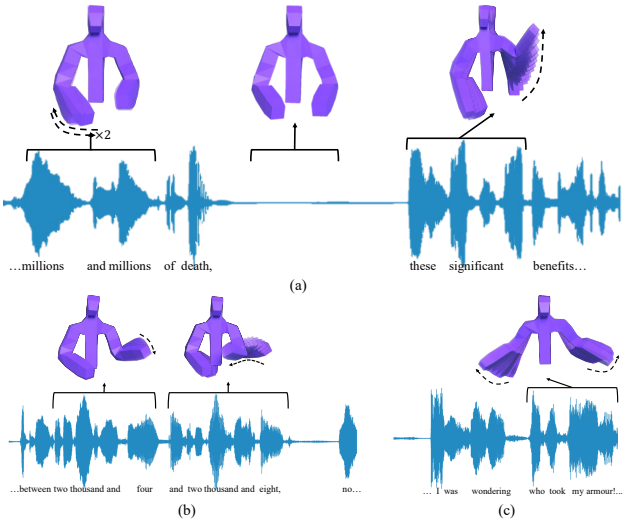

Figure 4: Sample results of general co-speech generation. The generated results show common characteristics based on the speech rhythm, emotion, context, and presence or absence of speech.

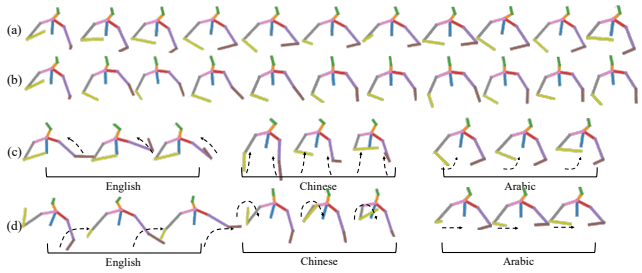

Figure 5: Sample results of cultural co-speech generation. (a) and (b) compare the generated gestures with Arabic and English. The results show Arabic prefers to use the right hand for more motions with the same multimodal inputs [15, 38]. (c) and (d) show examples of different cultures when speaking of "I" and future, respectively.

5.3 Quantitative Results

In this section, we use five quantitative evaluation metrics for comparison: L_1 distance, PCK, FGD, SRGR, and CDR. We also use user studies to evaluate the performance of generating gestures. Thus, a total of six evaluation metrics are used to better evaluate our method in different aspects. Additionally, the detail definitions of these indicators are shown in the supplementary materials.

Percent of correct keypoints (PCK). Previous works mainly use L_1 regression loss of the different models in comparison. Yang and Ramanan [72] propose PCK as a widely accepted metric for pose detection. The main idea is to calculate the proportion of the predicted keypoints within the α maximum pixel of the ground-truth keypoints. Following Ginosar et al. [21] suggestion, the α is set to 0.2 in our experiments. We test our method and seven baselines in Tab. 2. Our method outperforms other methods and increases the state-of-the-art PCK by 10.1% (from 70.6 to 77.7).

Fréchet gesture distance (FGD). Yoon et al. [73] propose FGD to apply the concept of FID [27] to the gesture generation problem. Accordingly, they trained a feature extractor based on autoencoding [55], which can be trained in an unsupervised manner. After that, FGD measures the latent features' distance between the generated gestures and real human gestures. We adopt FGD as the measure of gesture quality, the smaller value the better. Our method improves the average FGD score from 53.7 to 48.0 across three large-scale benchmark datasets (Tab. 2). Meanwhile, we also show the learning curve comparisons of FGD, which are shown in Fig. 6 (b). Compared with other baselines, our method achieves state-of-the-art performance with faster convergence speed.

Semantic-relevant gesture recall (SRGR). SRGR [42] is to evaluate the semantic relevancy of gestures, which can also be interpreted as whether the gestures are vivid and diverse. SRGR utilizes the semantic scores as a weight for the PCK between the generated gestures and the ground truth gestures. Compared with L_1 and PCK, SRGR emphasizes recalling gestures in the clip of interest and is more in line with the subjective human perception of gesture's valid diversity. As shown in Tab. 3, our approach achieves the highest score and increases the SRGR from 0.239 to 0.248.

Culture Deception Rate (CDR). CDR calculates the rate at which the network classifies the generated gestures into the correct cultural features, inspired by deception rate [56] from style transfer [20]. The definition of CDR is introduced in Sec. 4.5 and the results are shown in Tab. 3. Our method achieves the highest score with a cultural self-adapt architecture, which no need to train a new network for a different culture.

User study. Due to the gesture generation task being a highly subjective task, user study is widely adopted in the previous works [21, 42, 46, 60, 73]. Here we use two user evaluation metrics: gesture correctness and gesture-audio synchrony to evaluate the performance of each method. We select 20 audio-text pairs as the input of the above several compared methods, yielding 20 generated gestures for each method. We randomly ordered all the generated gestures and show them to participants and ask them two questions. The first is to evaluate gestures correctness, *i.e.*, physical correctness, diversity and attractive (more natural and human-like) [42]. The other is to choose the gesture motion which is more appropriate with the speech audio and words (gesture-audio synchrony) [73]. We collect 1,000 votes from 50 participants for each question and report the result in Fig. 6 (a), where we can see the results obtained by our method are more popular than those of other methods.

To summarize, our method can not only generate high-quality gestures based on audio and text, but also self-adapt to multiple cultures without training a new network. According to the user study, our method achieves both remarkable gestures correctness and gesture-audio synchrony.

5.4 Ablation Studies

In this section, we explore each component's effect on our method and validate their importance.

Text and audio encoder. Here, we compare the PCK and FGD values under different conditions in Tab. 4, including using different text and audio encoders, not fusing but only concatenating the features of audio and text, and the influence of different β_0 values. The results show that our method and architecture strike a balance by taking time cost and performance into account. For the β_0, when the demand for semantic relevancy is high, we encourage the network to generate gestures spatially similar to ground truth as much as possible, thus strengthening the L1 penalty and decreasing the adversarial penalty. Each component we used benefits for extracting the features of input modalities.

6 CONCLUSION

In this paper, to increase the cultural influence on co-speech gesture generation, we first build MCGD, a large freely available gesture dataset over 200 hours with multiple cultures, to solve the lack of datasets with different cultures. Furthermore, to overcome the generalization and flexibility across different languages, we propose a CSGN to synthesize gestures based on three modalities and adapt to multiple different cultures without requiring retraining the network. Statistical analyses on MCGD show that we balanced the datasets to cover a wide range of situations. It could benefit the research on comprehending relationships between multimodal data and developing naturalistic multimodal chatbots in different

Table 2: Comparison of PCK, FGD and L_1 distance score one three diverse datasets. The best scores are reported in bold.

Methods	PCK (↑)				FGD (↓)				L_1 distance (↓)			
	YouTube	TED	BEAT	Average	YouTube	TED	BEAT	Average	YouTube	TED	BEAT	Average
Seq2Seq	39.7	57.1	47.2	48.0	167.9	40.8	261.3	156.7	0.70	0.69	1.16	0.85
Speech2Gesture	54.5	67.2	51.8	57.8	105.4	30.1	256.7	130.7	0.67	0.64	1.13	0.81
Joint-Embedding	65.8	74.6	53.5	64.6	108.6	27.8	287.6	141.3	0.66	0.62	1.05	0.78
Audio2Gesture	56.7	72.9	54.8	61.5	78.6	21.7	223.8	108.0	0.61	0.59	1.05	0.75
MultiContext	67.4	75.3	56.4	66.4	40.5	6.2	176.2	74.3	0.56	0.57	0.98	0.70
ATG	67.0	78.3	58.1	67.8	37.1	6.2	156.4	66.6	0.57	0.53	0.92	0.67
CaMN	71.3	79.9	60.6	70.6	**31.6**	5.8	123.7	53.7	0.54	**0.52**	0.94	0.67
Ours	**80.7**	**86.5**	**65.9**	**77.7**	31.9	**4.9**	**107.2**	**48.0**	**0.54**	0.53	**0.91**	**0.66**

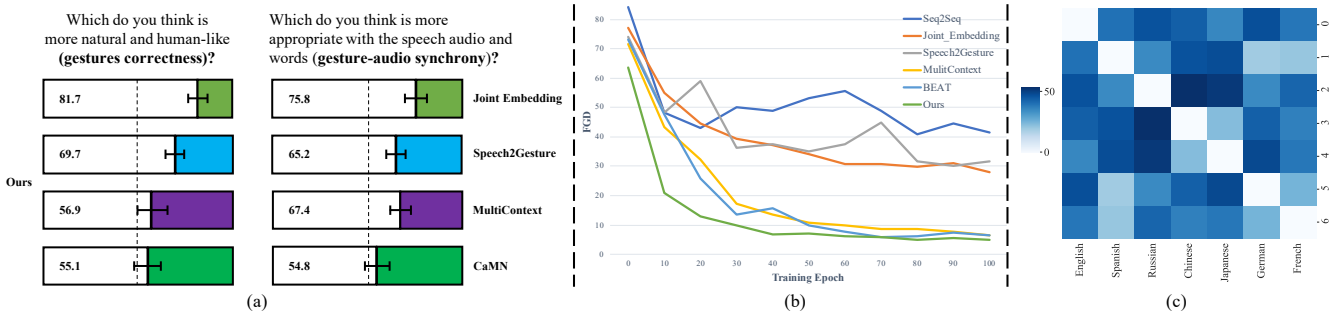

Figure 6: (a) Human perceptual study. We asked participants to choose which gesture gets better performance on "gesture correctness" or "gesture-audio synchrony". Our model is rated best for generating correct and audio-synchrony gestures. (b) Validation learning curves measured by Fréchet gesture distance (FGD). Our method achieves state-of-the-art performance with faster convergence speed. (c) Diverse distance of Z^C from different cultures. Color gradients correspond to Euclidean distance on a held-out test set (higher is better). Each cultural features are diverse from others and our CSGN is culture-specific.

Table 3: Results of SRGR (BEAT) and CDR (MCGD) comparisons with state-of-the-art models. Higher number indicates better performance. The best scores are reported in bold.

Methods	SRGR (↑)	CDR (↑)
Seq2Seq	0.173	12.80%
Speech2Gesture	0.092	18.41%
Joint-Embedding	0.127	24.62%
Audio2Gesture	0.097	21.85%
MultiContext	0.196	29.97%
ATG	0.205	28.95%
CaMN	0.239	33.63%
Ours	**0.248**	**39.87%**

cultures. Extensive experiments demonstrate our approach is concise and effective, and improves the state-of-the-art average FGD from 53.7 to 48.0 and culture deception rate from 33.63% to 39.87%.

ACKNOWLEDGMENTS

This work is supported by the Ng Teng Fong Charitable Foundation in the form of ZJU-SUTD IDEA Grant (188170-11102).

Table 4: The ablation studies on text and audio encoder.

Method		Train time sec/epoch (↓)	TED Gesture L_1 (↓)	TED Gesture FGD (↓)
β_0 (ours=0.4)	0.1	-	1.03	30.1
	0.3	-	0.85	10.3
	0.5	-	0.52	4.9
	0.7	-	0.51	8.9
TCN (ours=4)	3	338.7	0.57	10.2
	5	377.2	0.58	13.6
E_A	MultiContext	327.9	0.57	9.2
	ResNet50	635.8	0.54	6.3
	Wav2Clip	629.3	0.52	4.8
w/o fuse		353.6	0.53	5.4
Ours		359.2	0.53	4.9

REFERENCES

[1] Chaitanya Ahuja, Dong Won Lee, Yukiko I Nakano, and Louis-Philippe Morency. 2020. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 248–265.

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

[3] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 487–496.

[4] Simon Alexanderson, Éva Székely, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Generating coherent spontaneous speech and gesture from text. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–3.

[5] Ghazanfar Ali, Myungho Lee, and Jae-In Hwang. 2020. Automatic text-to-gesture rule generation for embodied conversational agents. *Computer Animation and Virtual Worlds* 31, 4-5 (2020), e1944.

[6] David F Armstrong, William C Stokoe, and Sherman E Wilcox. 1995. *Gesture and the nature of language*. Cambridge University Press.

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

[8] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. 2021. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE virtual reality and 3D user interfaces (VR)*. IEEE, 1–10.

[9] Lera Boroditsky. 2001. Does language shape thought?: Mandarin and English speakers' conceptions of time. *Cognitive psychology* 43, 1 (2001), 1–22.

[10] Paul Bremner, Anthony G Pipe, Chris Melhuish, Mike Fraser, and Sriram Subramanian. 2011. The effects of robot-performed co-verbal gesture on listener behaviour. In *2011 11th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 458–465.

[11] Alec Burmania, Srinivas Parthasarathy, and Carlos Busso. 2015. Increasing the reliability of crowdsourcing evaluations using online quality assessment. *IEEE Transactions on Affective Computing* 7, 4 (2015), 374–388.

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

[13] Justine Cassell, David McNeill, and Karl-Erik McCullough. 1999. Speech-gesture mismatches: Evidence for one underlying representation of linguistic and non-linguistic information. *Pragmatics & cognition* 7, 1 (1999), 1–34.

[14] Haibo Chen, Lei Zhao, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. 2021. Dualast: Dual style-learning networks for artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 872–881.

[15] Lena Darweesh. 2010. Nonverbal Behaviour as Communication: The Arabian Coffee Making Ritual. *International Journal of Interdisciplinary Social Sciences* 5, 7 (2010).

[16] Riccardo Del Gratta, Sara Goggi, Gabriella Pardelli, and Nicoletta Calzolari. 2021. Correction to: The LRE Map: what does it tell us about the last decade of our field? *Language Resources and Evaluation* 55 (2021), 285–286.

[17] Ylva Ferstl and Rachel McDonnell. 2018. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 93–98.

[18] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2020. Adversarial gesture generation with realistic gesture phasing. *Computers & Graphics* 89 (2020), 117–130.

[19] Anna Jia Gander, Nataliya Berbyuk Lindström, and Pierre Gander. 2021. Expressing Agreement in Swedish and Chinese: A Case Study of Communicative Feedback in First-Time Encounters. In *Cross-Cultural Design. Experience and Product Design Across Cultures: 13th International Conference, CCD 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part I 23*. Springer, 390–407.

[20] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2414–2423.

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

[22] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893* (2018).

[23] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. 2021. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*. 101–108.

[24] Edward T Hall. 1976. *Beyond culture*. Anchor.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[26] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. 2020. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–14.

[27] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[28] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[29] Peter J Huber. 1992. Robust estimation of a location parameter. *Breakthroughs in statistics: Methodology and distribution* (1992), 492–518.

[30] Ali Jahanian, Lucy Chai, and Phillip Isola. 2019. On the" steerability" of generative adversarial networks. *arXiv preprint arXiv:1907.07171* (2019).

[31] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* (2016).

[32] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759* (2016).

[33] Vinothini Kasinathan, Aida Mustapha, and Chow Khai Bin. 2021. A Customizable multilingual chatbot system for customer support. *Annals of Emerging Technologies in Computing (AETiC)* 5, 5 (2021), 51–59.

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

[35] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[36] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[37] Sotaro Kita. 2009. Cross-cultural variation of speech-accompanying gesture: A review. *Language and cognitive processes* 24, 2 (2009), 145–167.

[38] L Viola Kozak and Nozomi Tomita. 2012. On selected phonological patterns in Saudi Arabian Sign Language. *Sign Language Studies* 13, 1 (2012), 56–78.

[39] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.

[40] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S Srinivasa, and Yaser Sheikh. 2019. 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*. 763–772.

[41] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. 2021. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11293–11302.

[42] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022. BEAT: A Large-Scale Semantic and Emotional Multi-Modal Dataset for Conversational Gestures Synthesis. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*. Springer, 612–630.

[43] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. 2022. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10462–10472.

[44] Reza Lotfian and Carlos Busso. 2017. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing* 10, 4 (2017), 471–483.

[45] JinHong Lu, TianHang Liu, ShuZhuang Xu, and Hiroshi Shimodaira. 2021. Double-dcccae: Estimation of body gestures from speech waveform. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 900–904.

[46] Shuhong Lu and Andrew Feng. 2022. The DeepMotion entry to the GENEA Challenge 2022. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 790–796.

[47] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi.. In *Interspeech*, Vol. 2017. 498–502.

[48] Costanza Navarretta, Elisabeth Ahlsén, Jens Allwood, Kristiina Jokinen, and Patrizia Paggio. 2011. Creating comparable multimodal corpora for nordic languages. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*. 153–160.

[49] Rafael E Núñez and Eve Sweetser. 2006. With the future behind them: Convergent evidence from Aymara language and gesture in the crosslinguistic comparison of spatial construals of time. *Cognitive science* 30, 3 (2006), 401–450.

[50] Patrizia Paggio and Costanza Navarretta. 2011. Feedback and gestural behaviour in a conversational corpus of Danish. *NEALT (Northen European Association of*

Language Technology) Proceedings Series (2011), 33–39.

[51] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7753–7762.

[52] Vladimir Popescu, Lin Liu, Riccardo Del Gratta, Khalid Choukri, and Nicoletta Calzolari. 2016. New developments in the LRE map. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 4526–4530.

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

[54] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*. PMLR, 1278–1286.

[55] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1985. *Learning internal representations by error propagation*. Technical Report. California Univ San Diego La Jolla Inst for Cognitive Science.

[56] Artsiom Sanakoyeu, Dmytro Kotovenko, Sabine Lang, and Bjorn Ommer. 2018. A style-aware content loss for real-time hd style transfer. In *proceedings of the European conference on computer vision (ECCV)*. 698–714.

[57] Michael Studdert-Kennedy. 1994. Hand and Mind: What Gestures Reveal About Thought. *Language and Speech* 37, 2 (1994), 203–209.

[58] Kenta Takeuchi, Dai Hasegawa, Shinichi Shirakawa, Naoshi Kaneko, Hiroshi Sakuta, and Kazuhiko Sumi. 2017. Speech-to-gesture generation: A challenge in deep learning approach with bi-directional LSTM. In *Proceedings of the 5th International Conference on Human Agent Interaction*. 365–369.

[59] Emmi Toivio and Kristiina Jokinen. 2012. Multimodal Feedback Signaling in Finnish.. In *Baltic HLT*. 247–255.

[60] Daniela Trotta and Raffaele Guarasci. 2021. How are gestures used by politicians? A multimodal co-gesture analysis. *IJCoL. Italian Journal of Computational Linguistics* 7, 7-1, 2 (2021), 45–66.

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

[62] Andrea Vidal, Ali Salman, Wei-Cheng Lin, and Carlos Busso. 2020. MSP-face corpus: a natural audiovisual emotional database. In *Proceedings of the 2020 international conference on multimodal interaction*. 397–405.

[63] Ekaterina Volkova, Stephan De La Rosa, Heinrich H Bülthoff, and Betty Mohler. 2014. The MPI emotional body expressions database for narrative scenarios. *PloS one* 9, 12 (2014), e113647.

[64] Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview. *Speech Communication* 57 (2014), 209–232. https://doi.org/10.1016/j.specom.2013.09.008

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

[66] Bowen Wu, Chaoran Liu, Carlos Toshinori Ishi, and Hiroshi Ishiguro. 2021. Modeling the conditional distribution of co-speech upper body gesture jointly using conditional-GAN and unrolled-GAN. *Electronics* 10, 3 (2021), 228.

[67] Bowen Wu, Chaoran Liu, Carlos T Ishi, and Hiroshi Ishiguro. 2021. Probabilistic human-like gesture synthesis from speech using GRU-based WGAN. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*. 194–201.

[68] Jingyu Wu, Shi Chen, Wei Xiang, Lingyun Sun, Hongzeng Zhang, Zhenyu Zhang, and Yanxu Li. 2023. CNAMD corpus: A Chinese Natural Audiovisual Multimodal Database of Conversations for Social Interactive Agent. *International Journal of Human–Computer Interaction* (2023). https://doi.org/10.1080/10447318.2023.2228530

[69] Jingyu Wu, Lefan Hou, Zejian Li, Jun Liao, Li Liu, and Lingyun Sun. 2023. Preserving Structural Consistency in Arbitrary Artist and Artwork Style Transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 2830–2838.

[70] Elizabeth Würtz. 2005. Intercultural communication on web sites: A cross-cultural analysis of web sites from high-context cultures and low-context cultures. *Journal of computer-mediated communication* 11, 1 (2005), 274–299.

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

[72] Yi Yang and Deva Ramanan. 2012. Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence* 35, 12 (2012), 2878–2890.

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

[74] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. 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)*. IEEE, 4303–4309.

[75] Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037* (2023).

A ADDITIONAL EXPERIMENTS

A.1 Ablation Studies

Cultural self-adaptive module. We explore further experiments on the cultural self-adaptive module. We use different values of α and compare the performance with and without the average pooling layer in Tab. 5. We also calculate the diverse distance of $\mathbf{Z}^M$ from different cultures in Fig. 6 (c). Both results show the effectiveness of our method.

Table 5: The ablation study on cultural self-adaptive module

Method		MCGD CDR (↑)	Train time sec/epoch (↓)
α (ours=0.5)	0.0	35.27%	350.9
	0.3	37.67%	-
	0.7	39.56%	-
	1.0	38.11%	348.1
w/o average pooling		38.93%	351.0
Ours		39.87%	359.2

Each modality. Table. 6 summarizes the results of the ablation study, indicating that removing any of the three modalities - text, audio, and culture - led to a reduction in the model's performance. These results demonstrate that all three modalities used in the proposed model have positive effects on gesture generation.

Table 6: Results of the ablation study for the proposed model. Ablations are not accumulated

Configuration	FGD (↓)
Proposed (no ablation)	4.9
Without speech text modality	6.0
Without speech audio modality	6.1
Without speech culture modality	8.3

A.2 Additional Qualitative Results

Due to the fact that the performance of co-speech gesture generation cannot be well evaluated by images, in our supplementary materials (ZIP), we provide a video version of the results in the main paper, which can be found in a file named "Videos-Main". Additionally, to further demonstrate the effectiveness of our method, we also include more qualitative results in another file named "Videos-Supplementary".

B LIMITATION AND FUTURE WORK

While the current research presents promising results, there is still room for improvement. One major challenge is the difficulty of controlling the gesture generation process. Despite the possibility of manipulating cultural style, users cannot constrain gestures or control specific gestures, such as the desire for an avatar to make a deictic gesture when uttering a particular word. For example, we try to improve the controllability of the gesture generation process, the users can constrain gestures or control specific gestures, such as the desire for an avatar to make a deictic gesture when uttering a particular word. This lack of control is prevalent in most end-to-end neural network models, as pointed out by Jahanian et al. [30]. To address this issue, one potential solution would be to augment the existing model with more controllability features, such as constraining poses during generation or some rules based on prior knowledge.

In addition, it is worth noting that our current research focused on the motion of the upper body. However, integrating whole-body motion, including facial expressions and finger movements, would enhance the naturalness of the generated gestures. Meanwhile, more than 60% of scenes are from speech. We will add more data from other various scenes to ensure a balanced distribution of scenes in our future work.

Furthermore, our gesture generation process is based on Python, and the resulting gestures are animated in Blender using Python scripts. However, there are several software options available for constructing and manipulating 3D models. Although we have not yet explored how to utilize our generated gestures to manipulate models in these alternative software environments, it represents an interesting area for future work.

C TRAINING DETAILS

In the training stage, instead of taking a sequence of human poses as gestures, we represent each pose as directional vectors which represent the relative positions of the child joints from the parent joints and are favored for training [73]. In total, we have nine directional vectors: spine-neck, neck-nose, nose-head, neck-R/L shoulders, R/L shoulders-R/L elbows, and R/L elbows-R/L wrists. Each vector has three values indicating the 3-dimensions.

Meanwhile, due to the different FPS of video and frequencies of audio in each dataset, we uniformly resample the videos and audio at 15 FPS and 16K HZ.

At each training iteration, we randomly sample a batch of size N, each training sample has 34 frames which are sampled with a stride of 10 from the valid video sections. The initial four frames are used as seed poses and the model is trained to generate the remaining 30 poses (2 seconds). We excluded non-informative samples having little motion (i.e., low variance of a sequence of poses) and erratic samples having lying poses (i.e., low angle of the spineśneck vector).

We adopt Adam [35] with 0.005 learning rate, $\beta_1 = 0.5$ and $\beta_2 = 0.999$. Weights for the loss terms ($\beta_0\mathcal{L}_{rec}+\beta_1\mathcal{L}_{GAN,adv}+\beta_2\mathcal{L}_{cls}$) are determined experimentally ($\beta_0 = 500, \beta_1 = 5, \beta_2 = 3$). In addition, there is a warm-up period of 10 epochs in which the adversarial loss is not used ($\beta_1 = 0$). Note that in the ablation study of the main paper, the $\beta_0 = 0.5$ represents 0.5×10^3 for short. We train our model on eight 16 GB GeForce RTX 3090 GPUs, where batch size is 256 for training and epoch is 100 in total.

As for the comparison baselines, we use the publicly available implementations with the default configuration. It is worth noting that when we evaluate the CDR metric, we train other baselines for ten cultures because they need different networks for different cultures or languages.

Table 7: Additional MCGD statistical analysis.

Statistical Measure		Value
File Format	Video	Mp4
	Transcript	Txt
	Text Align	TextGrid
	Aduio	Wav
	Gesture Annotation	Npy
	Cache	Lmdb
Basci indicator	Resolution	720P
	Sampling frequency	25 HZ
	Bitrate	750 kbit/s
Gneder Distribution	Total perople	263
	Male	144
	Female	119

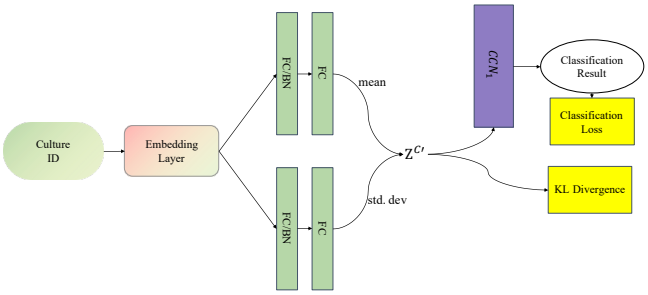

Figure 8: Detailed architecture of the cultural embedding space. The CCN_1 is the pre-trained model in the training stage of the main structure.

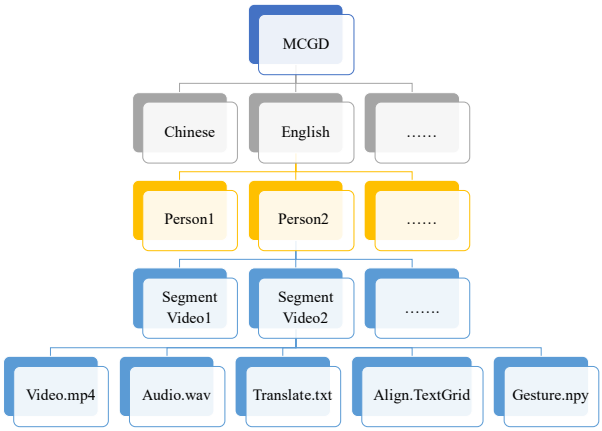

Figure 7: The file organization structure of MCGD.

D ADDTIONAL INFORMATION OF MCGD

In this section, we provide additional details about our MCGD and present further fundamental statistical analyses. The videos of our MCGD are obtained from official online websites, such as TED official channels and YouTube. Outdated videos of low resolution and videos of music performances or interviews are excluded. All videos are sourced from public domain videos or under Creative Commons licenses, which allow for such use. For the few videos without having transcripts, we use a multilingual ASR system USM [75], a single large model that performs ASR across 100+ languages and results in WER of 11.8%. The gesture we annotated contains a sequence of human poses p_s consisting of 10 upper body joints (spine, head, nose, neck, L/R shoulders, L/R elbows, and L/R wrists). All poses are spine-centered. In the representation of joint coordinates, a small translation of neck, which is the parent joint of both arms, can have an excessive effect on all coordinates of the arms. We denote human poses represented as directional vectors by d_i, and all directional vectors are normalized to the unit length. We note that forearm twists are not considered in this paper.

We also give additional statistical analysis and file organization structure of MCGD on Tab. 7 and Fig. 7, respectively.

E DETAILS OF CULTURAL STYLE EMBEDDING SPACE

The culture embedding space aims to convert cultural style to another or a specific style. For instance, if the input text/audio is in English, but the user wants the gestures to be in the Chinese style, the culture embedding space can be used. It utilizes the culture IDs to reflect the characteristics of each culture in the dataset. The culture IDs are represented as one-hot vectors where only one element of a selected speaker is nonzero.

A set of fully connected layers maps a speaker ID to a cultural embedding space $\mathbf{Z}^{C'} \in \mathbb{R}^{64}$ with the same dimensions as the E_C output $\mathbf{Z}^{C} \in \mathbb{R}^{64}$. To make the cultural embedding space more interpretable, variational inference [36, 54] that uses a probabilistic sampling process is used. In the training stage, as shown in Fig. 8, we take several steps to ensure that the cultural feature $\mathbf{Z}^{C'}$ is represented accurately and avoids trivial solutions. Firstly, we classify $\mathbf{Z}^{C'}$ using CCN_1. The CCN_1 has been pre-trained during the main structure training. This step helps in maintaining the correct decomposition of cultural features. Additionally, to prevent the cultural embedding space from becoming too sparse, we use the Kullback-Leibler (KL) divergence between $\mathcal{N}(0, I)$ and the cultural embedding space assumed Gaussian[36]. This step helps in preserving the diversity of the cultural features.

Moreover, our culture embedding space does differ from personal ID branches in previous works like SDT [53] and HA2G [43], which only use KL loss to ensure diversity. We also use gesture style loss Eq. 11 to train the culture ID branch. Furthermore, the culture ID branch in our methodology is not trained initially, unlike in SDT and HA2G. This process takes several steps to accurately represent cultural features and prevent trivial solutions. It's worth noting that within the learned culture embedding space, there might be some individual variations, yet they remain within the bounds of a particular cultural ID.

In total, the cultural embedding space has two loss functions: (1) a classification loss and (2) Kullbackś-Leibler (KL) divergence. Notice that the $\mathbf{Z}^{C'}$ on the style embedding space is trained separately from the main structure, and is only used in the synthesis stage.