EmotionGesture: Audio-Driven Diverse Emotional Co-Speech 3D Gesture Generation

Xingqun Qi*, Chen Liu*, Lingcheng Li, Jie Hou, Haoran Xin, Xin Yu†

***Abstract*—Generating vivid and diverse 3D co-speech gestures is crucial for various applications in animating virtual avatars. While most existing methods can generate gestures from audio directly, they usually overlook that *emotion* is one of the key factors of authentic co-speech gesture generation. In this work, we propose *EmotionGesture*, a novel framework for synthesizing vivid and diverse emotional co-speech 3D gestures from audio. Considering emotion is often entangled with the rhythmic beat in speech audio, we first develop an Emotion-Beat Mining module (EBM) to extract the emotion and audio beat features as well as model their correlation via a transcript-based visual-rhythm alignment. Then, we propose an initial pose based Spatial-Temporal Prompter (STP) to generate future gestures from the given initial poses. STP effectively models the spatial-temporal correlations between the initial poses and the future gestures, thus producing the spatial-temporal coherent pose prompt. Once we obtain pose prompts, emotion, and audio beat features, we will generate 3D co-speech gestures through a transformer architecture. However, considering the poses of existing datasets often contain jittering effects, this would lead to generating unstable gestures. To address this issue, we propose an effective objective function, dubbed Motion-Smooth Loss. Specifically, we model motion offset to compensate for jittering ground-truth by forcing gestures to be smooth. Last, we present an emotion-conditioned VAE to sample emotion features, enabling us to generate diverse emotional results. Extensive experiments demonstrate that our framework outperforms the state-of-the-art, achieving vivid and diverse emotional co-speech 3D gestures. Our code and dataset will be released at: *EmotionGestures*.**

***Index Terms*—Emotion Extraction, Diverse Co-speech Gesture, 3D Postures, Temporal Smooth**

I. Introduction

CO-SPEECH 3D gesture generation aims to synthesize vivid and diverse human poses consistent with the corresponding audio. This non-verbal body language helps people express their views and ideas more comprehensibly in daily communication [1]–[3]. Thus, animating virtual avatars with audio-coherent human gestures is crucial for various applications in embodied AI agents [4]–[8]. Conventionally, recent researchers build the end-to-end mapping between the speech audio and corresponding upper body dynamics [9]–[13]. They usually leverage a few initial postures as reference prompts to guide the generation [14].

Most existing co-speech gesture generation methods address this challenging task by constructing a large corpus and then modeling the correlation between audio and gestures [12], [13], [15], [16]. A few pioneering researchers have modeled emotional co-speech gestures by simply adding the emotion as a one-hot condition [17], [18]. These works overlook effectively exploring the emotional information of audio signals, resulting in unrealistic gestures in most real-world scenarios. Moreover, since the pose annotations of existing datasets [10], [19], [20] are often obtained by pre-trained 3D pose estimators, their poses usually contain jittering effects. Directly regressing postures from ground-truth might not yield smooth synthetic gesture sequences, as in previous studies [9]–[11].

As evidenced in previous works, there are three main challenges in this task:

- *How to effectively model the diverse emotional co-speech gestures?*
- *How to enforce the co-speech gestures to be well aligned with audio beats?*
- *How to achieve posture spatial-temporal smoothness when ground-truth 3D gestures are jittery?*

In this work, we propose a novel framework, ***EmotionGesture***, to generate vivid and diverse emotional 3D co-speech gestures driven by audio. In our EmotionGesture, we first propose an Emotion-Beat Mining (EBM) module to model emotional co-speech gestures. EBM extracts the emotion and audio beat features from the input audio signals. To achieve the rhythmic beats, previous works [9], [16] leverage the audio onset as an indicator. However, onset extraction may be affected by audio noise, thus producing low-fidelity audio-driven gestures. We observe that utter words with frame-wise timestamps can be employed to align the beats for the extracted audio beat features from BEM, as depicted in Figure 1. Hence, we design a contrastive learning fashion to enforce the beat features to be frame-wise aligned with audio rhythm through the synchronized transcripts. For emotion feature extraction, we employ an emotion classifier to ensure the extracted feature can represent the emotion in the audio.

Then, we propose an initial pose based Spatial-Temporal Prompter (STP) to generate future poses upon the initial poses. STP aims to ensure smoothness between the initial postures and future poses via prompt enhancement. Due to the mismatched temporal dimension of initial poses and target ones, previous works [10], [11], [20] straightly pad the temporal dimension of initial postures as reference prompt,

* these authors contributed equally to this work. This work was done when Xingqun Qi was an intern at the NetEase Fuxi AI Lab, Hangzhou, China.

† corresponding author: xin.yu@uq.edu.au

Xingqun Qi is with the Academy of Interdisciplinary Studies, The Hong Kong University of Science and Technology, Hong Kong, China. (e-mail: xingqun.qi@connect.ust.hk).

Chen Liu and Xin Yu are with the School of Electrical Engineering and Computer Science, The University of Queensland, Queensland, Australia. (e-mail: uqcliu32@uq.edu.au, xin.yu@uq.edu.au)

Lingcheng Li, Jie Hou, and Haoran Xin are with the NetEase Fuxi AI Lab, Hangzhou, China. (e-mail: lilincheng@corp.netease.com, houjie1@corp.netease.com, xinhaoran@corp.netease.com)

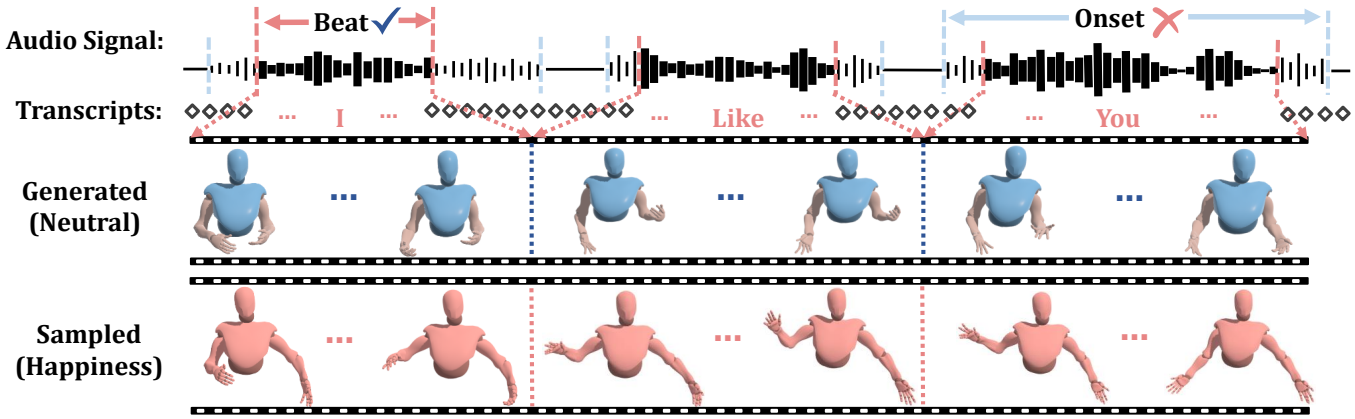

Fig. 1. Diverse **emotional** exemplary clips sampled by our ***EmotionGesture***. We identify the beat via frame-wise aligned utter words (pink) in audio-synchronized transcripts. Due to the noisy environment, it is improper to directly extract audio onsets (blue) as rhythmic indicators.

which might result in unnatural and ambiguous gestures. Instead of this, our STP fully takes advantage of the motion clues from the initial poses to produce future postures via two prompt learning strategies: spatial-interpolation prompt learning and temporal-reinforcement prompt learning. The spatial-interpolation prompt learning strategy leverages the embedding of initial postures as guidance to update each future frame via a learnable interpolation manner. It provides spatial-wise smoothness but fails in the long-term sequential constraints. As a complement, the temporal-reinforcement prompt learning strategy aggregates historical temporal changes to consolidate the temporal correlation for future sequential steps. Afterward, we concatenate the initial pose features and future pose features as the enhanced pose prompt to guide the smooth gesture generation.

Furthermore, considering the jittering issue of pose annotations, we design a simple yet effective objective function, dubbed Motion-Smooth Loss, inspired by soft label smoothing technique [21], [22]. Here, we produce the smoothed motion offset of the ground truth through a temperature coefficient. Then, the smoothed motion offset is applied to provide supervision of generating temporally smoothed gestures. In addition, we introduce an emotion-conditioned VAE [23] to sample the diverse emotion features based on these emotion-clustered features. In this fashion, our framework enables diverse emotional gesture generation, as displayed in Figure 1.

Moreover, the current largest emotional co-speech gesture dataset [17] only includes 30 avatar identities with limited pre-defined talking topics. This may lead to insufficient diversity of 3D postures and speech content. Therefore, we newly collected a TED Emotion dataset composed of more than 1.7K avatar identities from TED talk show videos. Extensive experiments on BEAT and TED Emotion datasets demonstrated that our ***EmotionGesture*** significantly outperforms various counterparts, displaying vivid and diverse emotional co-speech 3D gestures.

To summarize, our main contributions are four-fold:

- We devise EmotionGesture, a novel framework that achieves audio-driven diverse emotional co-speech 3D gesture generation.
- We propose an Emotion-Beat Mining (EBM) module to facilitate diverse emotional gesture generation while aligning generated gestures with audio beats.
- We present a Spatial-Temporal Prompter (STP) to obtain the enhanced temporal-coherency pose prompt, thus guiding the smooth gesture generation.
- We design a simple yet effective Motion-Smooth Loss to overcome the pose jittering issues in existing datasets, thus achieving temporally smooth co-speech gestures.

II. Related Work

A. Co-speech Gesture Generation.

Co-speech gesture generation aims at synthesizing audio-synced human pose sequences of the talking people. It has witnessed impressive research interests for its wide practical value in various applications like human-agent interaction [24]–[27], robotics [19], [28], and holoportation [29]. Thus, numerous studies have been proposed to tackle this challenging task. Traditionally, previous researchers follow the rule-based pipelines that the mapping between speech and gestures is pre-defined by linguistic experts [4], [30]–[32]. In this pattern, researchers focus on refining the transitions matching process between the generated different motions. However, it may need expensive efforts for the experts when facing complex scenes.

Recently, thanks to more and more released datasets [10], [17], [19], [20], co-speech gesture generation is significantly improved by deep-learning based approaches. Among various approaches, many researchers intend to utilize multi-modality clues to build the associations between co-speech gestures and audio signals [33], text transcripts [34], and speaker identity [10], [20]. Only a few research [14], [17] simply explore the significance between generated gestures and emotions. However, they fail to fully exploit the emotional pheromones of speech audio, resulting in lower emotion-related and unrealistic gestures in most real-world scenarios. In this work, we propose to extract the diverse emotion representation from input speech audio, thus achieving audio-driven emotion control of the generated gestures.

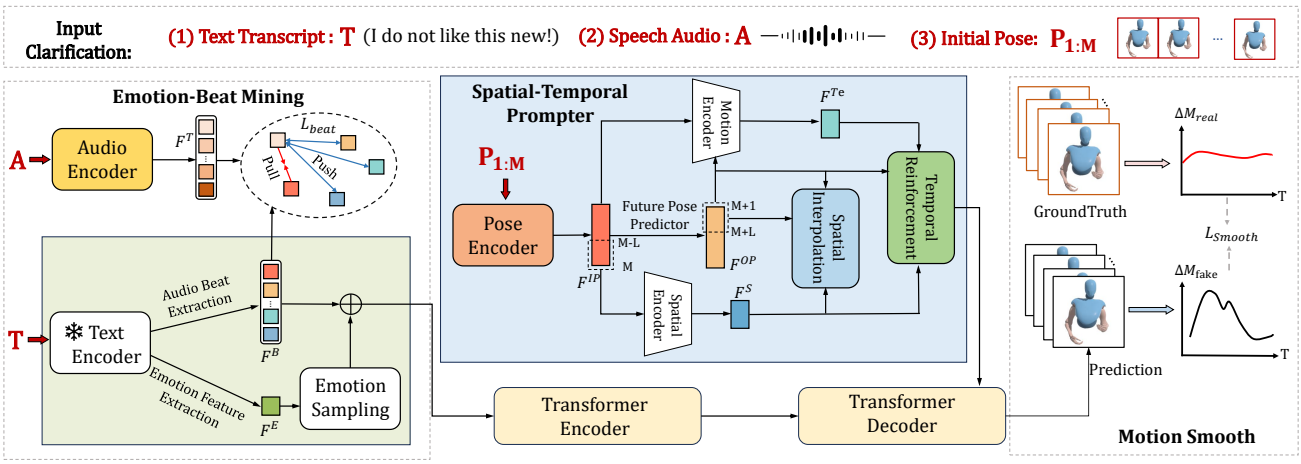

Fig. 2. Overview of our proposed ***EmotionGesture*** framework. With extracted audio beat features F^B, and emotion features F^E, we could achieve the generation of audio-driven diverse emotional co-speech gestures. Our spatial-temporal prompter aims to obtain the enhanced temporal-coherency pose prompt based on the initial pose sequence.

B. 3D Human Motion Modeling.

Modeling 3D human motion plays a critical role in research areas of both computer vision and graphics. One of the key targets is to preserve the spatial-temporal coherency of the generated human motions [35]–[38]. To achieve that, previous studies [39]–[42] directly employ the discrete cosine transform (DCT) as the post-processing strategy. However, these straightforward approaches fail in the jittering problem in pseudo-annotated 3D co-speech gesture datasets. Motivated by these methods, we propose a spatial-temporal prompter to keep the smoothness between the initial poses and generated gestures. In addition, we design a simple yet effective motion-smooth loss function to smooth our generated co-speech motions. Notably, our motion-smooth loss gives a practical solution to address the jittering issue. Such design could prospectively provide insights into relevant domains not only on co-speech gesture generation but also on 3D human pose estimation [43], motion prediction [44], and talking head synthesis [45].

III. Proposed Method

A. Problem Formulation

Given speech audio sequence $A = \{a_1, ..., a_N\}$ as input, the goal of our framework $\mathcal{G}$ is to generate continuous diverse emotional 3D co-speech poses as $P = \{p_1, ..., p_N\}$, where N denotes the total frame number corresponding to A. Here, p_i represents J joints of the human upper body including two hands. To ensure that the generated co-speech gestures are diverse emotional while preserving alignment with the audio beat, we also introduce the emotion label and text transcripts $T = \{t_1, ..., t_N\}$ to provide supervision in the training phase. Notice that the emotion label actually is the one-hot vector in our framework. With the aforementioned representations, the overall objective is expressed as:

$$\underset{\mathcal{G}}{argmin} \left\| P - \mathcal{G}\left(A, \{p_1, ..., p_M\}\right) \right\|, \tag{1}$$

where $\{p_1, ..., p_M\}$ is initial pose sequence.

B. Emotion-Beat Mining Module

To achieve the audio-driven emotion control on the generated diverse co-speech gesture and maintain the rhythmic alignment with speech audio, the inherently entangled emotion features and rhythmic beat have to be extracted independently from the input speech audio. Thus, we propose an Emotion-Beat Mining (EBM) module to extract the features of emotion and audio beat. As illustrated in Figure 2, we leverage an audio encoder E_a combined with two MLP-based projectors to obtain two separate features, *i.e.*, audio beat features $F^B = \{f_1^B, ..., f_N^B\}$ and emotion features F^E.

1) Beat Alignment: We intend to enforce the embedded beat features being frame-wise aligned with the speech rhythm. Intuitively, previous works utilize the signal processing technique [46] directly identify the onsets of audio signals as the speech rhythm [9], [16]. Instead of this, our beat-alignment strategy is built upon the insight that "***beat starts when people are speaking***". Thus, we introduce the text transcripts synchronized with frame-wise timestamps (*i.e.*, each uttered word is aligned with a frame-wise gesture) to provide beat alignment supervision via contrastive learning.

In the audio-synchronized test transcripts, the pause duration is inserted with padding tokens ($\diamond$) while the uttered words are encoded as unique identifiers. As shown in Figure 2, we leverage a pre-trained word2vec model [47] as our text encoder E_t to obtain the audio-coherency transcript features $F^T = \{f_1^T, ..., f_N^T\}$. In our contrastive learning formulation, we utilize the uttered word feature f_u^T as the anchor sample, and only the beat feature aligned with the current uttered word serves as a positive sample, denoted as f_u^{B+}. Then, the negative samples are defined as other timesteps beat features in the same sequence. In this manner, beat features of other timesteps that reflect different rhythmic and semantic information are repelled. Drawing inspiration from InfoNCE [48], our beat-

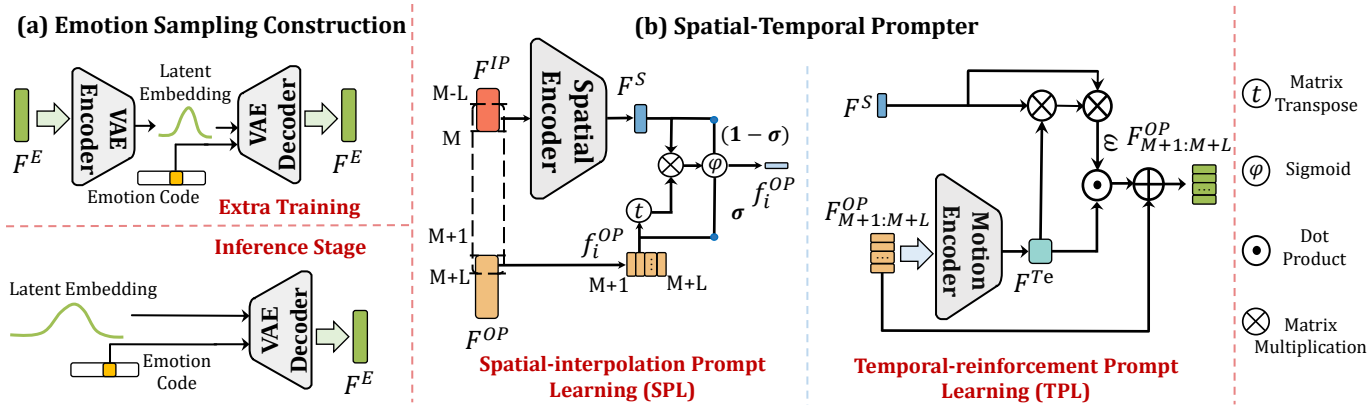

Fig. 3. Details of our proposed emotion sampling component and spatial-temporal prompter.

alignment contrastive learning loss is expressed as:

$$\mathcal{L}_{beat} = \mathbb{I}_{[U\neq 0]}\left(-log\sum_{u=1}^{U}\frac{exp\left(sim\left(f_u^T, f_u^{B+}\right)/\tau\right)}{\sum_{i=1,i\neq u}^{N-1}exp\left(sim\left(f_u^T, f_i^{B-}\right)/\tau\right)}\right), \tag{2}$$

where U is the total number of uttered words, $\mathbb{I}_{[U\neq 0]} \in [0,1]$ is an indicator function evaluating to 0 if there are no utterances in this audio duration. $sim(\cdot)$ denotes cosine similarity, and τ is the temperature hyperparameter. Different from the HA2G model [10] directly taking the sequence-wise words features F^T to constrain the audio embedding features F^B keeping semantic consistency, globally. Our contrastive learning loss aims to achieve frame-wise beat alignment. Meanwhile, we effectively prevent model collapse when there are no uttered words in the audio duration (*i.e.*, no rhythmic onsets).

2) Emotion Sampling: In order to fully extract the emotional features, we leverage the emotion label to provide the classification supervision via a classification header. The supervision is formulated as the cross-entropy loss: $\mathcal{L}_{emo} = -\sum_{c=1}^{C} y_c \cdot log\left(q_c\right)$, where C represents the number of emotions, y_c denotes whether the sample belongs to emotion label c, and q_c is the prediction probability. Afterward, to achieve the diverse emotion gesture generation, once our EmotionGesture is well trained, we freeze the parameters of the overall framework and train an emotion-conditioned VAE [23] model upon clustered-emotion features. Once we obtain sampled diverse emotion features, we feed the summation of emotion features and beat features into the transformer-based backbone to produce co-speech 3D gestures. In the inference phase, the emotion-conditioned VAE samples diverse emotion features to produce diverse emotional gestures, as depicted in Figure 3.

C. Spatial-Temporal Prompter

Conventionally, researchers directly take the padded initial poses as the conditioned prompt to guide the co-speech gesture synthesis. Unlike the previous works [10], [11], [20], we aim to build the enhanced temporal-coherency pose prompt based on the initial pose sequence. This pose prompt would lead to smooth co-speech gesture generation. Thus, we propose a Spatial-Temporal Prompter for keeping smoothness among the enhanced pose prompt. As shown in Figure 2, we first utilize a pose encoder E_p to obtain the initial pose embedding $F^{IP} = \{f_1^{IP}, ..., f_M^{IP}\}$. Then, we produce the future pose features $F^{OP} = \{f_{M+1}^{OP}, ..., f_N^{OP}\}$ via a 1-D convolution-based pose predictor. Inspired by [38], we nominate a transition chunk consisting of the last L frames of initial pose embeddings and the first L ones of the future pose features. Here, the dimension of each frame in the transition chunk is $\mathbb{R}^{1\times D}$. STP keeps the transition from the first L poses to the last L poses to be a smooth sequence. Concretely, STP consists of two prompt learning strategies, *i.e.*, spatial-interpolation prompt learning and temporal-reinforcement prompt learning.

1) Spatial-interpolation Prompt Learning.: As shown in Figure 3, our spatial-interpolation prompt learning strategy aims to ensure spatial-wise smoothness between the first L frames and the last L frames in a learnable interpolation pattern. Concretely, we first design a spatial encoder to obtain the ensembled spatial representation of the first L poses, donated as $F^S \in \mathbb{R}^{1\times D}$. Then, we calculate the spatial interpolation score between the ensembled spatial representation and each frame in the last L poses. With the help of the interpolation score, each spatial smoothed feature of the last L frames is represented as:

$$\begin{aligned} f_i^{OP} &= \sigma f_i^{OP} + (1-\sigma)F^S, \\ \sigma &= \varphi\left(F^S \otimes (f_i^{OP})'\right), \\ i &\in [M+1, M+L], \end{aligned} \tag{3}$$

where σ is the interpolation score, φ is the sigmoid operation, $\otimes$ is matrix multiplication, $'$ indicates the transpose operation.

2) Temporal-reinforcement Prompt Learning.: To further enhance the long-term spatial-temporal smoothness in the transition chunk, we propose a temporal-reinforcement prompt learning strategy. Specifically, we develop a motion encoder to obtain the sequence-aware temporal embedding of the last L pose features, denoted as $F^{Te} \in \mathbb{R}^{L\times 1}$. Here, the temporal embedding represents the temporal changes among predicted L poses. Then, we compute the long-term spatial-temporal correlation score between the ensembled spatial representation and temporal embedding, globally. Once we obtain this long-

TABLE I
THE ACCURACY OF OUR PRE-TRAINED EMOTION CLASSIFIER ON THE BEAT DATASET.

Emotion	Neutral	Anger	Happiness	Fear	Disgust	Sadness	Contempt	Suprise	Average
Accuracy(%)	99.91	98.80	99.84	99.60	100.00	99.67	99.75	98.30	**99.70**

term correlation score, the temporal-reinforced last L poses sequence in transition is attained by:

$$F^{OP}_{M+1:M+L} = F^{OP}_{M+1:M+L} + F^{OP}_{M+1:M+L} \odot \omega, \\ \omega = \mathcal{S}\left((F^{Te} \otimes F^{S}) \otimes F^{S}\right), \tag{4}$$

where $\odot$ denotes dot product, $\mathcal{S}$ means softmax operation, $\omega \in \mathbb{R}^{L\times 1}$ is the calculated long-term spatial-temporal correlation score. Afterward, we concatenate the initial pose embedding F^{IP} and future pose features F^{OP} as the enhanced pose prompt.

To fully take advantage of the initial poses, we leverage the enhanced pose prompt as the query Q. Then, we use Q to match the key features K and value features V in the transformer-based decoder via three times Multi-Head Attention (MHA) [49], expressed as:

$$MultiHead(Q, K, V) = softmax(\frac{QK}{\sqrt{d}})V. \tag{5}$$

In this fashion, sequence-aware correspondence between the emotional audio representation and pose prompt is jointly built. Then, similar to [50], we employ a motion discriminator to ensure the generated co-speech gestures preserve realism.

D. Training Objectives

1) Motion-Smooth Loss.: To address the jittering problem in most existing co-speech 3D gesture datasets, we design a simple yet effective objective, named Motion-Smooth Loss. Our motion-smooth loss aims to produce the smoothed motion offset as the target. Concretely, we first compute the motion offset of the jittering ground truth as $\triangle\mathcal{M}_{real} \in \mathbb{R}^{(N-1)\times D}$. Meanwhile, we obtain the motion offset of the generated gestures as $\triangle\mathcal{M}_{fake} \in \mathbb{R}^{(N-1)\times D}$ in the same way. Inspired by soft label smoothing technique [21], [22], we leverage a smooth temperature coefficient to produce the smoothed ground truth. Then, our motion-smooth loss is formulated as:

$$\mathcal{L}_{smooth} = \mathcal{KL}\left(\mathcal{S}\left(\triangle\mathcal{M}_{real}/\Gamma\right) || \, \mathcal{S}\left(\triangle\mathcal{M}_{fake}\right)\right), \tag{6}$$

where Γ is the smooth temperature coefficient, $\mathcal{S}$ means softmax operation, $\mathcal{KL}$ denotes Kullback Leibler Divergence.

2) Reconstruction Loss.: We leverage the ground truth gestures to constrain the generated co-speech gestures as:

$$\mathcal{L}_{rec} = \left\| P - \tilde{P} \right\|_1, \tag{7}$$

where $\tilde{P}$ denotes generated gestures.

3) Adversarial Learning.: Following the configuration of previous works, we employ the adversarial training loss as:

$$\mathcal{L}_{adv} = \mathbb{E}_P\left[log\mathcal{D}(P)\right] + \mathbb{E}_A\left[log(1 - (\mathcal{G}\left(A, \{p_1, ..., p_M\}\right))\right], \tag{8}$$

where $\mathcal{D}$ denotes the discriminator and $\mathcal{G}$ means generator. Finally, the **overall objective** is:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \mathcal{L}_{total} = \lambda_r \mathcal{L}_{rec} + \mathcal{L}_{adv} + \lambda_b \mathcal{L}_{beat} \\ + \lambda_e \mathcal{L}_{emo} + \lambda_s \mathcal{L}_{smooth}, \tag{9}$$

where $\mathcal{L}_{beat}$ denotes our beat-alignment contrastive learning loss and $\mathcal{L}_{emo}$ indicates our emotion classification loss. The λ_r, λ_b, λ_e, and λ_s are weight coefficients.

IV. EXPERIMENTS

A. Datasets and Experimental Setting

1) BEAT Dataset.: BEAT [17] is a large-scale multi-emotion dataset for conversational gestures synthesis, including audio, transcripts, and 3D human whole-body motions from 30 speakers. In BEAT, the audio and corresponding gestures are annotated with 8 emotional styles (*i.e.*, *neutral, anger, happiness, fear, disgust, sadness, contempt, and surprise*) in 4 languages. In our experiments, similar to [10], [19], [20], we use the upper body with 47 joints of English speakers, amounting to about 35 hours duration. Meanwhile, we resample the human motion with 15 FPS and select the continuous 60 frames with a stride of 30 as gesture clips. Finally, we obtain 55,420 clips. The training, validation, and testing subsets are divided following the proportion as 70%, 10%, and 20%. In particular, the numbers of sequences in each data partition are Training set: 38,814; Validation set: 5,502; Testing set: 11,10. Similar to [17], we keep the partition of different emotions in each sub-dataset as: Neutral 51%, Anger 7%, Happiness 7%, Fear 7%, Disgust 7%, Sadness 7%, Contempt 7%, and Suprise 7%.

2) TED Emotion Dataset.: Inspired by [10], [12], we newly collect a TED Emotion dataset based on in-the-wild TED talk show videos. In particular, we leverage the state-of-the-art human pose estimator ExPose [52] to obtain the 43 3D upper body joints as pseudo ground truth. Then we keep a similar data processing strategy with the BEAT dataset to acquire 107,468 gesture clips, and each clip has 60 frames. To acquire the emotion label of the TED Emotion dataset, we leverage the BEAT to pre-train an audio-based emotion classifier. Concretely, we leverage the BEAT training set to pre-train the classifier, and the validation set to select the optimal model. Finally, the emotion classification accuracy on the BEAT testing set is 99.70%. For each emotion, the accuracy is reported in Table I.

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE BEAT DATASET AND TED EMOTION DATASET OF THE GENERATED CO-SPEECH GESTURES. ↓ INDICATES THE LOWER THE BETTER, AND ↑ INDICATES THE HIGHER THE BETTER. ± MEANS 95% CONFIDENCE INTERVAL. † MEANS OUR EMOTIONGESTURE FRAMEWORK IS IMPLIED WITHOUT THE SAMPLING PHASE DURING INFERENCE.

Settings	Models	BEAT Dataset						TED Emotion Dataset					
		L2↓	MPJRE↓	FGD↓	BA↑	EA(%)↑	Diversity↑	L2↓	MPJRE↓	FGD↓	BA↑	EA(%)↑	Diversity↑
Constant Gestures	Seq2Seq [19]	2.48	5.90	1.11	0.30	65.24	-	2.10	5.70	1.63	0.50	66.43	-
	S2G [51]	2.14	3.82	1.09	0.79	65.39	-	1.68	4.70	0.65	0.79	68.10	-
	JointEM [34]	2.33	4.08	2.05	0.28	53.01	-	1.24	3.38	0.73	0.26	58.45	-
	CAMN [17]	1.97	3.56	1.12	0.74	71.96	-	1.24	3.27	0.87	0.87	71.86	-
	Ours †	**1.59**	**2.69**	**0.47**	**0.93**	**81.21**	**-**	**0.93**	**2.37**	**0.11**	**0.93**	**85.59**	**-**
Diverse Gestures	Trimodal [20]	-	-	2.94	0.86	32.31	$30.52^{\pm 0.41}$	-	-	0.76	0.77	69.35	$16.68^{\pm 0.48}$
	HA2G [10]	-	-	2.45	0.76	57.10	$11.90^{\pm 0.47}$	-	-	0.71	0.75	72.30	$9.10^{\pm 0.32}$
	DiffGes [11]	-	-	2.03	0.82	15.62	$36.61^{\pm 0.61}$	-	-	1.28	0.82	33.41	$18.28^{\pm 2.42}$
	TalkShow [12]	-	-	0.91	0.84	45.01	$20.57^{\pm 0.70}$	-	-	0.83	0.84	74.20	$11.42^{\pm 0.92}$
	Ours	**-**	**-**	**0.52**	**0.91**	**81.16**	**39.34**$^{\pm 1.02}$	**-**	**-**	**0.13**	**0.91**	**84.81**	**19.46**$^{\pm 0.19}$

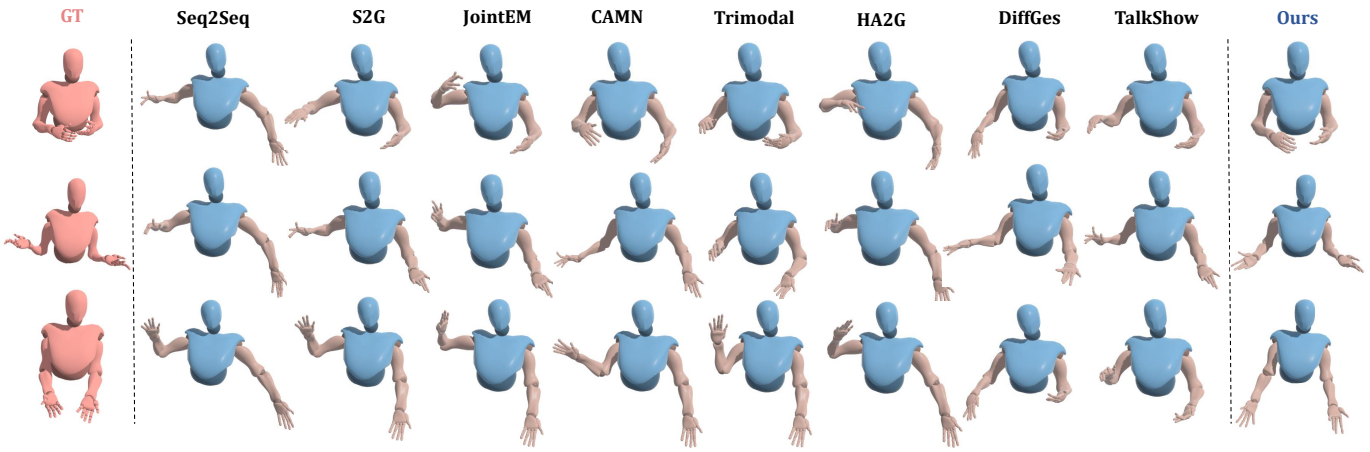

Fig. 4. Visualization of our predicted 3D hand gestures against various state-of-the-art methods [10]–[12], [17], [19], [20], [34], [51]. From top to bottom, we show the three keyframes (an early, a middle, and a late one) of a pose sequence. Best view on screen.

Then, we annotate the emotion label of the TED Emotion dataset via this emotion classifier. To ensure the annotation accuracy of emotion categories, we randomly visualize gestures for each emotion category and set the classification threshold ≥ 0.9. Meanwhile, we drop the two uncommon emotions, *disgust and contempt*. Finally, we totally gain 78,734 gesture clips. Our experiments adopt the same division criteria as [10] to split the training, validation, and testing subsets. The numbers of sequences in each data partition of our newly annotated TED Emotion dataset are Training set: 66,679; Validation set: 6,258; Testing set: 5,797. Our newly collected TED Emotion Dataset will be released at *EmotionGestures*.

3) Implementation Details.: We set the whole length of the clip as $N = 60$, and the length of initial poses as $M = 10$ for both datasets. Similar to [10], [16], we employ ResNetSE-34 [53] as our audio encoder. We adopt three stacking blocks and leverage the 2D-convolution-based header to map the dimension of audio features to be $N \times 512$, where N is the temporal dimension. Then, we employ two MLP-based projectors to extract the emotion features and audio beat features. For initial poses, we propose an MLP-based pose encoder E_p to obtain the embedded initial pose features. Then, we leverage a 1D convolution-based pose predictor to produce the future pose sequence.

For both BEAT and TED Emotion datasets, the frequency of speech audio is pre-processed to 16kHZ. Then, for more compact signal information preservation, the audio signals are converted to mel-spectrograms with the FFT window size is 1024 and hop length is 512. Finally, the audio signals are represented as the 2D time-frequency mel-spectrogram of size 128×124. Similar to [50], [54], we convert the original 3D joints of both datasets into the 6D rotation representations [55]. The 6D rotation representation has proved effective for training neural networks due to its continuity.

We set feature dimension $D = 512$, and $L = 10$. Empirically, we set $\lambda_r = 100$, $\lambda_b = 0.05$, $\lambda_e = 0.1$, $\lambda_s = 0.5$, $\tau = 0.1$ in Eq. (2), and $\Gamma = 10$ in Eq. (6). Our model is implemented on the PyTorch platform with 2 NVIDIA RTX 2080Ti GPUs. We adopt the Adam optimizer with an initial learning rate of 0.0002. The whole training takes 100 epochs with a batch size of 128.

TABLE III
ABLATION STUDY ON DIFFERENT LOSS FUNCTIONS AND DIFFERENT COMPONENTS OF OUR PROPOSED EMOTIONGESTURE FRAMEWORK. ↓ INDICATES THE LOWER THE BETTER, AND ↑ INDICATES THE HIGHER THE BETTER. ± MEANS 95% CONFIDENCE INTERVAL. ‡ DENOTES WE DIRECTLY CONCATENATE THE ONE-HOT EMOTION LABEL WITH AUDIO FEATURES AS INPUT. "SPATIAL" MEANS THE SPATIAL-INTERPOLATION PROMPT LEARNING, AND "TEMPORAL" DENOTES THE TEMPORAL-REINFORCEMENT PROMPT LEARNING. "+" INDICATES THAT WE CONTINUE TO ADD THE CORRESPONDING COMPONENT OR LOSS FUNCTION UPON "BASELINE", SEQUENTIALLY. NOTICE THAT ONLY WITH THE SAMPLING SETTING OUR FRAMEWORK COULD ACHIEVE DIVERSE CO-SPEECH GESTURE SYNTHESIS.

Ablation	BEAT Dataset						TED Emotion Dataset					
Settings	L2↓	MPJRE↓	FGD↓	BA↑	EA(%)↑	Diversity↑	L2↓	MPJRE↓	FGD ↓	BA↑	EA(%)↑	Diversity↑
Baseline	1.99	3.52	1.16	0.90	59.43	-	1.36	3.47	1.29	0.90	59.59	-
Baseline‡	1.97	3.52	1.11	0.91	68.73	-	1.35	3.46	1.29	0.90	64.46	-
+ EAD	1.94	3.44	0.98	0.91	61.84	-	1.30	3.21	0.92	0.91	66.13	-
+ $\mathcal{L}_{emo}$	1.88	3.32	0.83	0.91	72.35	-	1.25	2.95	0.79	0.91	76.62	-
+ $\mathcal{L}_{beat}$	1.76	3.03	0.67	0.92	75.90	-	1.05	2.78	0.52	0.92	81.60	-
+ Spatial	1.73	2.97	0.61	0.92	78.89	-	0.99	2.60	0.40	0.92	82.91	-
+ Temporal	1.68	2.89	0.55	0.93	80.42	-	0.94	2.41	0.30	0.92	85.35	-
+ $\mathcal{L}_{smooth}$	**1.59**	**2.69**	**0.47**	**0.93**	**81.21**	**-**	**0.93**	**2.37**	**0.11**	**0.93**	**85.59**	**-**
+ Sampling	**-**	**-**	**0.52**	**0.91**	**81.16**	**39.34**$^{\pm 1.02}$	**-**	**-**	**0.13**	**0.91**	**84.81**	**19.46**$^{\pm 0.19}$

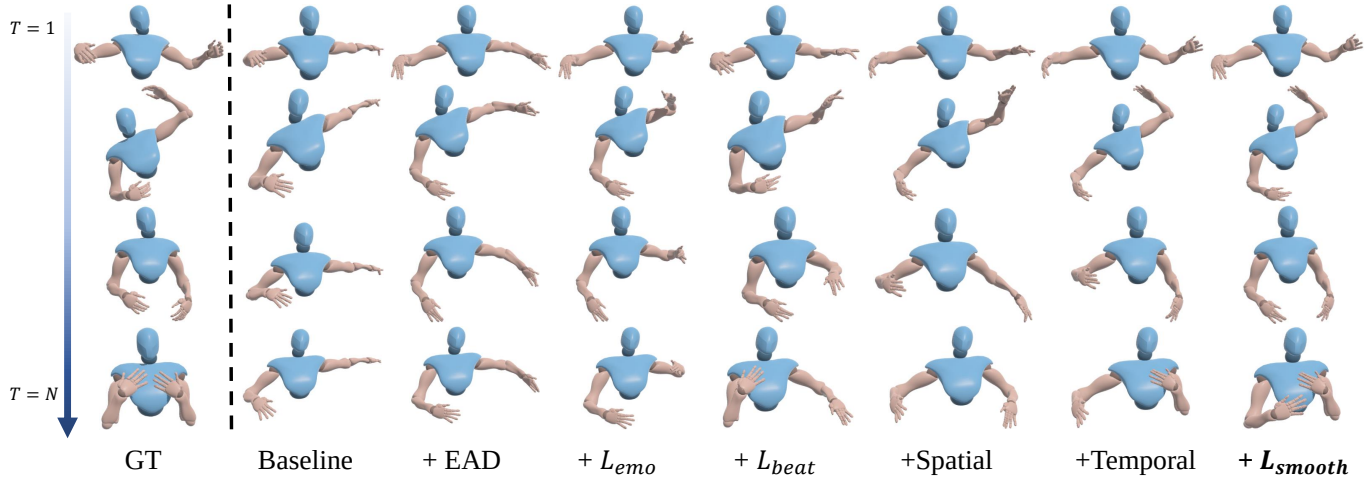

Fig. 5. Visual comparisons of ablation study. We show the key frames of the generated gestures. From top to bottom, we show four key frames (an early, two middle, and a late one) of a pose sequence. Best view on screen.

4) Evaluation Metrics.: To fully evaluate the superior performance of our proposed EmotionGesture framework, we employ the following metrics:

- $L2$ **Distance**: Distance between the generated co-speech gestures and target ones.
- **MPJRE**: Mean Per Joint Rotation Error [○] (MPJRE) [54], [56] measures the absolute distance between the systhesized 3D representation joints and the pseudo ground truth.
- **FGD**: Fréchet Gesture Distance (FGD) [20] evaluates the distribution distance between the ground truth and synthesized gesture. We pre-train an autoencoder based on BEAT and TED Emotion datasets to compute this metric.
- **BA**: Beat Alignment (BA) score [17] measures the rhythmic matching degree between the generated gestures and input audio.
- **EA**: Emotion Accuracy (EA) score measures whether the generated co-speech gestures represent the same emotion with audio. It is calculated by our pre-trained gesture-based emotion classifier.
- **Diversity**: Diversity score [15], [57] indicates the difference among the generated gestures.

B. Quantitative Results

1) Comparisons with the state-of-the-art.: We compare our EmotionGesture with previous state-of-the-art co-speech gesture generation approaches in recent years. We observe that the existing research could be roughly divided into two categories: 1) Directly generated the constant results from input audio signals, *i.e.*, Seq2Seq [19], S2G [51], JointEM [34], and CAMN [17]. 2) Generate the diverse co-speech gestures upon audio signals, *i.e.*, Trimodal [20], HA2G [10], DiffGes [11], and TalkShow [12]. Our framework can achieve optimal results in both experimental settings. For fair comparisons, all the models are implemented by source codes or pre-trained models released by authors. The experimental results

TABLE IV
ABLATION STUDY ON THE INFLUENCE OF TRANSITION CHUNK LENGTH L IN THE SPATIAL-TEMPORAL PROMPTER. ↓ INDICATES THE LOWER THE BETTER, AND ↑ INDICATES THE HIGHER THE BETTER. "DUPLICATE" INDICATES REPEATING THE INITIAL POSES TO ACHIEVE THE SAME TEMPORAL DIMENSION AS THE TARGET ONES. SIMILARLY, "ZERO" MEANS DIRECTLY PADDING THE TEMPORAL DIMENSION OF INITIAL POSES WITH ZERO ELEMENTS. NOTICE THAT IN THE PADDING STRATEGY, WE DROP THE SPATIAL-TEMPORAL PROMPTER IN THE EXPERIMENTS.

Variant Model	Setting	BEAT Dataset				
		L2 ↓	MPJRE ↓	FGD ↓	BA ↑	EA(%) ↑
Padding	Zero	2.11	3.49	1.00	0.21	65.38
	Duplicate	1.96	3.39	0.78	0.88	73.78
Chunk Length	$L = 2$	1.69	2.91	0.57	0.91	76.10
	$L = 4$	1.62	2.78	0.54	0.91	76.56
	$L = 6$	1.61	2.76	0.52	0.92	78.13
	$L = 8$	1.60	2.69	0.49	0.92	78.88
	$L = 10$	**1.59**	**2.69**	**0.47**	**0.93**	**81.21**

TABLE V
THE USER STUDY ON NATURALNESS, SMOOTHNESS, AND AUDIO-GESTURE SYNCHRONY. THE RATING SCORE RANGE IS 1-5, WITH 5 BEING THE BEST. ↑ INDICATES THE HIGHER THE BETTER.

Methods	Seq2Seq [19]	S2G [51]	JointEM [34]	CAMN [17]	Trimodal [20]	HA2G [10]	DiffGes [11]	TalkShow [12]	**Ours**
Naturalness↑	2.92	2.33	2.83	3.38	3.98	3.87	4.43	4.30	**4.56**
Smoothness↑	2.91	2.63	2.74	3.11	3.39	2.89	3.71	4.45	**4.75**
Synchrony↑	3.07	2.64	2.63	3.30	3.02	3.55	3.80	4.72	**4.80**

of the counterparts are re-implemented under the same setting as ours. The comparisons are divided into two parts.

First, we adopt the $L2$ Distance, MPJRE, FGD, BA, and EA for a well-rounded view of the constant generation comparisons. As reported in Table II, our framework outperforms all the competitors with a large marginal gap. For instance, we surpass all methods on the metrics FGD, BA, and EA with both two datasets. Remarkably, on the BEAT dataset, our framework is even 56.9% (*i.e.*, $(1.09 - 0.47)/1.09 \approx 56.9\%$) lower than the sub-optimal method on the FGD metric. This indicates our synthesized co-speech gestures are much more realistic than other counterparts. Meanwhile, the highest scores of BA and EA demonstrate our emotion-beat miming module facilitates the generated gestures to be frame-wise rhythmic while preserving emotion control. As for the diversity result comparison, we exploit the FGD, BA, EA, and Diversity metrics to verify the superior performance of our framework. Our framework exceeds all the counterparts markedly from the perspective of diversity. Although the FGD, BA, and EA are slightly worse than the constant experimental setting due to the sampled emotion features, our framework still achieves optimal results. Meanwhile, we observe that the diversity scores of DiffGes [11] are much closer to ours due to being well adapted by the diffusion model [58]. However, the huge inference time makes this model demonstrate poor practical values in the real-time co-speech gesture applications.

2) Ablation Study.: We conduct the ablation study to demonstrate the effectiveness of each proposed component and loss function, as displayed in Table III. Our baseline model is implemented by a simple transformer-based encoder-decoder backbone without emotion-beat mining in two branches. "+ EAD" in Table III means we only adopt the two MLP-based projectors without other constraints. As shown in Table III, all the combinations of our proposed modules and loss functions have positive impacts on the co-speech gesture generation. Specifically, even if we directly adopt the two MLP-based projectors without any other auxiliary losses (*i.e.*, "+ EAD"), our framework reaches competitive results than utilizing the one-hot emotion vector (*i.e.*, "Baseline‡"). After we add the emotion classification loss $\mathcal{L}_{emo}$, the EA score is significantly improved on both datasets, Then, adopting the beat-alignment loss $\mathcal{L}_{beat}$ ideally improves the performance on $L2$ distance, MPJRE, FGD, and BA metrics. This supports our insight on "onsets start when people are speaking".

To verify the effectiveness of our proposed STP, we split it into two steps in the ablation. First, we just leverage the spatial-interpolation prompt learning strategy, our framework achieves the lower $L2$ distance, MPJRE, and higher EA. Then, we employ the temporal-reinforcement prompt learning to reach better performance, especially on the metrics of $L2$ distance, MPJRE. Both of these two components encourage the generated gestures to be more consistent with initial poses. Next, we adopt our simple yet effective motion-smooth loss $\mathcal{L}_{smooth}$. As noticed in Table III, the FGD metric on both datasets decreased drastically. This indicates that $\mathcal{L}_{smooth}$ ensures the generated gestures realize realistic smooth gesture distribution.

Additionally, to fully verify the effectiveness of the spatial-temporal prompter, we conduct ablation experiments on transition chunk length L. For fair comparisons, the experiments are implemented on the BEAT dataset without sampling during the inference phase, as reported in Table IV. The padding strategy achieves the much lower performance than our proposed spatial-temporal prompter, especially on the BA score.

This suggests that directly padding initial poses as the pose prompt would lead to the generated gestures being low-fidelity and mismatched with the audio rhythmic beat. Although our prompt enhancement strategy is slightly influenced by the chunk length, it still significantly surpasses the direct padding upon the same length initial pose sequence. Even if our chunk length is 2, we still achieve better performance.

C. Qualitative Evaluation

1) User Study.: Moreover, we conduct the user study based on the recruited 15 volunteers to better analyze the visual quality of the generated co-speech 3D gestures by various methods. Inspired by [10], [11], [15], our user study adopts *naturalness, smoothness, and synchrony* as the main evaluation perceptions with the rating score range being 1-5 (the higher, the better). As displayed in Table V, the average statistical results show that our framework gains the best performance in all three metrics. Especially in the *naturalness* and *synchrony*, our framework achieves significant advantages over all the other counterparts. This strongly proves the effectiveness of our proposed beat-alignment strategy and motion-smooth loss.

2) Visualization.: To fully verify the performance of our framework, we show the keyframes visualization of generated co-speech 3D gestures among our results against all the competitors in Figure 4. We observe that the Seq2Seq, S3G, and JointEM generate unnatural co-speech gestures compared with the ground truth. CAMN, Trimodal, and HA2G synthesize the natural gestures while their gestures are misaligned with the audio rhythm. In addition, although the DiffGes and TalkShow can generate better gestures than others, they stiffly match the audio signal by swinging the arms. Meanwhile, the gestures generated by these methods are less emotion-coherency with the audio. On the contrary, our framework reaches vivid and diverse co-speech gesture generation. As depicted in Figure 4 and Figure 1, our framework enables the natural and diverse emotional gesture synthesis (*i.e.*, neutral-to-happiness, the gestures swing at a wider angle).

Moreover, to demonstrate the effectiveness of our proposed different components and loss functions, we visualize the keyframes of the generated co-speech gestures. As illustrated in Figure 5, we can clearly observe that each component and loss function have a positive impact on the visualization of generated gestures.

V. Conclusion

In this paper, we propose a novel framework ***EmotionGesture*** to generate audio-driven vivid and diverse emotional co-speech 3D gestures. To achieve the audio-driven emotion control of the generated gestures, we fully take advantage of the audio-coherency transcripts for obtaining the emotional audio representation via the emotion-beat mining module. Then, we propose a spatial-temporal prompter to maintain the smoothness of the generated co-speech gestures. Moreover, we design a simple yet effective motion-smooth loss to smooth the jittering movement of the generated results. Extensive experiments on the BEAT and our newly collected TED Emotion datasets demonstrate the superior performance of our work with competitive ones.

Our framework may produce some failure cases for some rarely-seen emotions (*e.g.*, the disgust and contempt emotional gestures may not be easy to distinguish). We will explore handling the rarely-seen emotional gestures with a better generalization ability model in the future. Considering the broader impact, the generated co-speech gestures may be misleveraged in malicious avatar forgery. However, we believe our proposed technique would facilitate the research on multi-modality learning in a proper way of the real applications.

References

[1] J. Cassell, D. McNeill, and K.-E. McCullough, "Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information," *Pragmatics & cognition*, vol. 7, no. 1, pp. 1–34, 1999.

[2] P. Wagner, Z. Malisz, and S. Kopp, "Gesture and speech in interaction: An overview," pp. 209–232, 2014.

[3] J. P. De Ruiter, A. Bangerter, and P. Dings, "The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis," *Topics in cognitive science*, vol. 4, no. 2, pp. 232–248, 2012.

[4] C.-M. Huang and B. Mutlu, "Robot behavior toolkit: generating effective social behaviors for robots," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012, pp. 25–32.

[5] S. Y. Gadre, K. Ehsani, S. Song, and R. Mottaghi, "Continuous scene representations for embodied ai," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 14 849–14 859.

[6] A. Padmakumar, J. Thomason, A. Shrivastava, P. Lange, A. Narayan-Chen, S. Gella, R. Piramuthu, G. Tur, and D. Hakkani-Tur, "Teach: Task-driven embodied agents that chat," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 2017–2025.

[7] P. Wolfert, N. Robinson, and T. Belpaeme, "A review of evaluation practices of gesture generation in embodied conversational agents," *IEEE Transactions on Human-Machine Systems*, 2022.

[8] R. Jahanmahin, S. Masoud, J. Rickli, and A. Djuric, "Human-robot interactions in manufacturing: A survey of human behavior modeling," *Robotics and Computer-Integrated Manufacturing*, vol. 78, p. 102404, 2022.

[9] Y. Liang, Q. Feng, L. Zhu, L. Hu, P. Pan, and Y. Yang, "Seeg: Semantic energized co-speech gesture generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 473–10 482.

[10] X. Liu, Q. Wu, H. Zhou, Y. Xu, R. Qian, X. Lin, X. Zhou, W. Wu, B. Dai, and B. Zhou, "Learning hierarchical cross-modal association for co-speech gesture generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 462–10 472.

[11] L. Zhu, X. Liu, X. Liu, R. Qian, Z. Liu, and L. Yu, "Taming diffusion models for audio-driven co-speech gesture generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[12] H. Yi, H. Liang, Y. Liu, Q. Cao, Y. Wen, T. Bolkart, D. Tao, and M. J. Black, "Generating holistic 3d human motion from speech," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[13] H. Liu, N. Iwamoto, Z. Zhu, Z. Li, Y. Zhou, E. Bozkurt, and B. Zheng, "Disco: Disentangled implicit content and rhythm learning for diverse co-speech gestures synthesis," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3764–3773.

[14] T. Ao, Z. Zhang, and L. Liu, "Gesturediffuclip: Gesture diffusion model with clip latents," *ACM Trans. Graph.*

[15] X. Liu, Q. Wu, H. Zhou, Y. Du, W. Wu, D. Lin, and Z. Liu, "Audio-driven co-speech gesture video generation," *arXiv preprint arXiv:2212.02350*, 2022.

[16] T. Ao, Q. Gao, Y. Lou, B. Chen, and L. Liu, "Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 6, pp. 1–19, 2022.

[17] H. Liu, Z. Zhu, N. Iwamoto, Y. Peng, Z. Li, Y. Zhou, E. Bozkurt, and B. Zheng, "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, 2022, pp. 612–630.

[18] S. Yang, Z. Wu, M. Li, Z. Zhang, L. Hao, W. Bao, M. Cheng, and L. Xiao, "Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models," in *Proceedings of the 32nd International Joint Conference on Artificial Intelligence, IJCAI 2023*. ijcai.org, 2023.

[19] Y. Yoon, W.-R. Ko, M. Jang, J. Lee, J. Kim, and G. Lee, "Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4303–4309.

[20] Y. Yoon, B. Cha, J.-H. Lee, M. Jang, J. Lee, J. Kim, and G. Lee, "Speech gesture generation from the trimodal context of text, audio, and speaker identity," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–16, 2020.

[21] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[22] Z. Allen-Zhu and Y. Li, "Towards understanding ensemble, knowledge distillation and self-distillation in deep learning," in *International Conference on Learning Representations*, 2023.

[23] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, 2015.

[24] H. S. Koppula and A. Saxena, "Anticipating human activities for reactive robotic response." in *IROS*. Tokyo, 2013, p. 2071.

[25] A. Huang, P. Knierim, F. Chiossi, L. L. Chuang, and R. Welsch, "Proxemics for human-agent interaction in augmented reality," in *CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–13.

[26] M. Salem, S. Kopp, I. Wachsmuth, K. Rohlfing, and F. Joublin, "Generation and evaluation of communicative robot gesture," *International Journal of Social Robotics*, vol. 4, pp. 201–217, 2012.

[27] Z. Wang, X. Qi, K. Yuan, and M. Sun, "Self-supervised correlation mining network for person image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7703–7712.

[28] C. T. Ishi, D. Machiyashiki, R. Mikata, and H. Ishiguro, "A speech-driven hand gesture generation method and evaluation in android robots," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3757–3764, 2018.

[29] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou *et al.*, "Holoportation: Virtual 3d teleportation in real-time," in *Proceedings of the 29th annual symposium on user interface software and technology*, 2016, pp. 741–754.

[30] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone, "Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents," in *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, 1994, pp. 413–420.

[31] S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro, "Virtual character performance from speech," in *Proceedings of the 12th ACM SIGGRAPH/Eurographics symposium on computer animation*, 2013, pp. 25–35.

[32] I. Poggi, C. Pelachaud, F. de Rosis, V. Carofiglio, and B. De Carolis, "Greta. a believable embodied conversational agent," *Multimodal intelligent information presentation*, pp. 3–25, 2005.

[33] J. Li, D. Kang, W. Pei, X. Zhe, Y. Zhang, Z. He, and L. Bao, "Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 293–11 302.

[34] C. Ahuja and L.-P. Morency, "Language2pose: Natural language grounded pose forecasting," in *2019 International Conference on 3D Vision (3DV)*. IEEE, 2019, pp. 719–728.

[35] N. Battan, Y. Agrawal, S. S. Rao, A. Goel, and A. Sharma, "Glocalnet: Class-aware long-term human motion synthesis," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 879–888.

[36] C. Zhong, L. Hu, Z. Zhang, Y. Ye, and S. Xia, "Spatio-temporal gating-adjacency gcn for human motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6447–6456.

[37] W. Mao, R. I. Hartley, M. Salzmann *et al.*, "Contact-aware human motion forecasting," *Advances in Neural Information Processing Systems*, vol. 35, pp. 7356–7367, 2022.

[38] W. Mao, M. Liu, and M. Salzmann, "Weakly-supervised action transition learning for stochastic human motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8151–8160.

[39] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade, "Nonrigid structure from motion in trajectory space," *Advances in neural information processing systems*, vol. 21, 2008.

[40] Y. Huang, F. Bogo, C. Lassner, A. Kanazawa, P. V. Gehler, J. Romero, I. Akhter, and M. J. Black, "Towards accurate marker-less human shape and pose estimation over time," in *2017 international conference on 3D vision (3DV)*. IEEE, 2017, pp. 421–430.

[41] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9489–9497.

[42] Q. Zhao, C. Zheng, M. Liu, P. Wang, and C. Chen, "Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[43] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3d human pose estimation with spatial and temporal transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 656–11 665.

[44] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano, "Human motion diffusion model," *arXiv preprint arXiv:2209.14916*, 2022.

[45] R. Huang, W. Zhong, and G. Li, "Audio-driven talking head generation with transformer and 3d morphable model," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 7035–7039.

[46] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on speech and audio processing*, vol. 13, no. 5, pp. 1035–1047, 2005.

[47] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[48] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[50] E. Ng, S. Ginosar, T. Darrell, and H. Joo, "Body2hands: Learning to infer 3d hands from conversational gesture body dynamics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 865–11 874.

[51] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik, "Learning individual styles of conversational gesture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3497–3506.

[52] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. J. Black, "Monocular expressive body regression through body-driven attention," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*. Springer, 2020, pp. 20–40.

[53] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," *arXiv preprint arXiv:2003.11982*, 2020.

[54] J. Jiang, P. Streli, H. Qiu, A. Fender, L. Laich, P. Snape, and C. Holz, "Avatarposer: Articulated full-body pose tracking from sparse motion sensing," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*. Springer, 2022, pp. 443–460.

[55] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5745–5753.

[56] X. Qi, C. Liu, M. Sun, L. Li, C. Fan, and X. Yu, "Diverse 3d hand gesture prediction from body dynamics by bilateral hand disentanglement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[57] H.-Y. Lee, X. Yang, M.-Y. Liu, T.-C. Wang, Y.-D. Lu, M.-H. Yang, and J. Kautz, "Dancing to music," *Advances in neural information processing systems*, vol. 32, 2019.

[58] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.