Weakly-Supervised Emotion Transition Learning for Diverse 3D Co-speech Gesture Generation

Xingqun Qi[1], Jiahao Pan[1], Peng Li[1], Ruibin Yuan[1], Xiaowei Chi[1], Mengfei Li[1]
Wenhan Luo[2], Wei Xue[1], Shanghang Zhang[3], Qifeng Liu[1,✉], Yike Guo[1,✉]
[1] The Hong Kong University of Science and Technology
[2] Sun Yat-sen University [3] Peking University
xingqun.qi@connect.ust.hk, {liuqifeng, yikeguo}@ust.hk

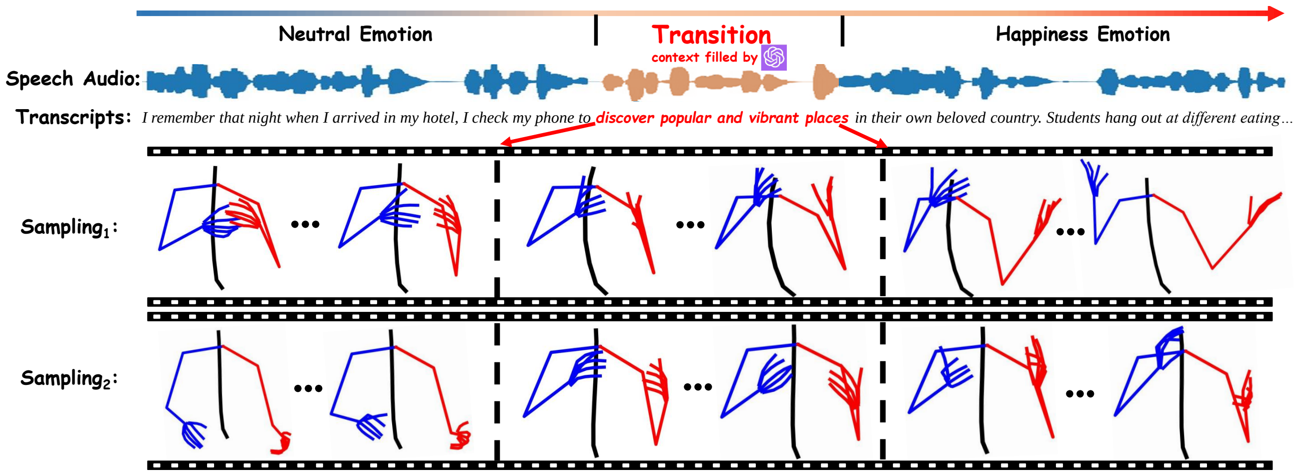
Neutral Emotion
Transition
context filled by
Happiness Emotion
Speech Audio:
Transcripts:
I remember that night when I arrived in my hotel, I check my phone to discover popular and vibrant places in their own beloved country. Students hang out at different eating…
Sampling₁:
Sampling₂:

Figure 1. Diverse exemplary clips sampled by our method from **our newly collected BEAT Emotion Transition Dataset**. The vital frames are visualized to demonstrate that the upper body gestures change with the emotion transition of human speech, synchronously. From top to bottom: the input speech audio, the corresponding transcript, and two sampled clips. Best view on screen.

Abstract

Generating vivid and emotional 3D co-speech gestures is crucial for virtual avatar animation in human-machine interaction applications. While the existing methods enable generating the gestures to follow a single emotion label, they overlook that long gesture sequence modeling with emotion transition is more practical in real scenes. In addition, the lack of large-scale available datasets with emotional transition speech and corresponding 3D human gestures also limits the addressing of this task. To fulfill this goal, we first incorporate the ChatGPT-4 and an audio inpainting approach to construct the high-fidelity emotion transition human speeches. Considering obtaining the realistic 3D pose annotations corresponding to the dynamically inpainted emotion transition audio is extremely difficult, we propose a novel weakly supervised training strategy to encourage authority gesture transitions. Specifically, to enhance the coordination of transition gestures w.r.t. different emotional ones, we model the temporal association representation between two different emotional gesture sequences as style guidance and infuse it into the transition generation. We further devise an emotion mixture mechanism that provides weak supervision based on a learnable mixed emotion label for transition gestures. Last, we present a keyframe sampler to supply effective initial posture cues in long sequences, enabling us to generate diverse gestures. Extensive experiments demonstrate that our method outperforms the state-of-the-art models constructed by adapting single emotion-conditioned counterparts on our newly defined emotion transition task and datasets.

1. Introduction

Co-speech gesture generation aims to synthesize vivid and emotional human postures coordinated with the audio input. These non-verbal behaviors serve as a key factor during human conversations that significantly facilitates the delivery of speech content [3, 12, 14]. Meanwhile, model-

✉ Corresponding authors.

ing co-speech gestures has a wide range of embodied AI applications in human-machine interaction [20, 25], robot assistants [7], and virtual/augmented reality (AR/VR) [9]. Conventionally, many researchers usually focus on synthesizing human upper body gestures consistent with speech audio [29, 44, 46, 49].

Nevertheless, except for a few recent works that generate the co-speech gestures of a single emotion category [2, 26, 48], previous works mostly focus on emotion-agnostic generation [44, 46]. Most of them overlook synthesizing the long sequence co-speech gestures with the emotion transitions, which are more practical in real-world scenes. For example, a person may not maintain a single emotion forever when communicating with others or in speech talking. In this work, we therefore introduce the task of ***speech-driven emotion transition*** for generating vivid and diverse 3D co-speech gestures, displayed in Figure 1. There are two main challenges in this task: 1) Datasets of 3D co-speech gestures synchronized with emotion transition speech audios are scarce. Creating such a dataset containing 3D human postures is difficult due to the lack of guidance from emotional experts and complex motion capture systems. 2) Modeling the plausible and temporal coherent co-speech gestures from one emotion to another in long sequences is difficult, especially in the transition duration.

To overcome the issue of data scarcity, we newly present two datasets containing emotion-transition human speech that are built upon previous single-emotion ones [26, 29]. In particular, thanks to the developed language model ChatGPT-4 [31], we first leverage it to generate text transcripts of the transition based on speech context. Then, by employing the audioLDM2 technique [27, 28] for audio inpainting, we ensure the inpainted transition's timbral consistency with its adjacent contexts and a smooth emotional transition throughout. Afterwards, to support our insight on modeling co-speech gestures coherent with emotion-transition speech, corresponding 3D human postures of transition are required.

However, due to the dynamically generated transition transcripts, it is infeasible to construct the aligned realistic 3D pose annotation of human bodies. Hence, we solve the challenges of vivid co-speech gesture generation in a novel weakly supervised pattern, containing a motion transition infusion mechanism and an emotion mixture strategy. Specifically, in the motion transition infusion mechanism, we model the temporal correlation between the generated head and tail gesture features as style guidance representation. The style guidance representation provides motion transition cues that are infused into the transition embedding via an adaptive instance normalization (AdaIN) layer [16]. In this manner, we can effectively enhance the coordination of transition gestures w.r.t. two different emotional ones.

Moreover, to alleviate the lack of supervision during the transition between two emotions, the emotion mixture strategy is built to provide weak emotional supervision of the generated transition gestures. Concretely, we learn a joint embedding of two different emotional gesture sequences using a temporal aggregation encoder. Then, we pre-train an emotion classifier based on the annotated human 3D poses with single emotion labels in the dataset. Here, this joint embedding is leveraged as an emotion mixture weight for the pre-trained classifier to facilitate high-fidelity transition gesture synthesis with desirable properties. Finally, considering the generated 3D postures should be non-deterministic given the human speech, we devise a keyframe sampler to produce diverse initial poses as reference. In this fashion, our method enables diverse co-speech gesture generation with emotion transitions. Extensive experiments conducted on our newly constructed two datasets verify the effectiveness of our methods, displaying vivid and emotional 3D co-speech gestures.
Overall, our contributions are summarized as follows:

- We introduce a new task of emotion transition co-speech gesture generation cooperating with two newly constructed datasets named BEAT Emotion Transition (BEAT-ETrans) and TED Emotion Transition (TED-ETrans), significantly facilitating research on 3D human motion modeling.
- We design a motion transition infusion mechanism to ensure the temporal coordination of transition gestures w.r.t. two different emotional ones and a weakly supervised emotion mixture strategy to enable high-fidelity transition gesture synthesis with desirable properties.
- Extensive experiments show that our method outperforms state-of-the-art counterparts on both datasets, displaying realistic and vivid co-speech gestures given emotion transition human audios.

2. Related Work

2.1. Co-speech Gesture Synthesis

Synthesizing human co-speech gestures plays a significant role in various applications [15, 20, 34, 43]. Numerous studies have been proposed to address these issues that are roughly divided into rule-based approaches [18, 19], machine-learning-based approaches [21, 38], and deep learning-based ones [2, 26, 29, 33, 44, 46, 48, 49]. Traditional researchers follow the rule-based workflow, leveraging the speech-gesture pairs as guidance to generate co-speech gestures pre-defined by linguistic experts. Other early works integrate the manually defined gesture features with machine learning techniques to synthesize the co-speech gestures. In the aforementioned two manners, the researchers usually focus on the optimization of the matching process between human speech and pre-defined gestures. It may require experts much expensive effort in the speech-gesture pair construction.

Recently works focus on building the mapping directly from the input human speech and sequential gestures by exploiting the deep neural networks. They usually leverage multi-modality cues to facilitate the generation of co-speech gestures, associating with the speech audio [23, 44, 49], speaker identity [46], emotion [2, 26], and text transcript [29]. However, they overlook that emotion transition of the long sequential co-speech gesture modeling is much more practical in the real scenes. Moreover, since the lack of annotated 3D gestures corresponding to dynamically constructed transition speech, few of the above methods could be directly adapted to this new thought.

2.2. 3D Human Motion Modeling

Human motion modeling aims to generate realistic and smooth human motions with various multi-modality conditions [6, 17, 30], including co-speech gesture generation as a sub-task. One of the hottest topics is synthesizing human motion from text prompts with a few past postures as the seed [32, 42, 47]. These methods usually engage in forcing human motion to represent the semantic expression aligned with the text. Literally, the task most closely related to ours is AI choreographer [22, 24, 39] which generates the motion from music signals. However, the AI choreographer works mainly focus on the rhythmic-coherent motion of the whole human body but without subtle finger gestures. While sharing a similar goal with the approaches mentioned above, our newly defined work differs from them significantly. We take the emotion-transition long sequence co-speech gesture without corresponding 3D human pose annotation into consideration, thus motivating us to utilize the motion transition of two annotated emotional gestures and coherent the overall sequence.

3. Proposed Method

3.1. Emotion Transition Dataset Construction

We aim to address the emotion transition co-speech gesture generation in a weakly supervised manner. Due to the existing paired speech-gesture datasets [26, 33], we could focus on synthesizing the high-fidelity transition human audios. Synthesizing datasets conducive to our task focuses on ensuring semantic coherence, smooth emotional audio transitions, consistent timbre, and audio fidelity.

Preliminary: Considering our key insight to modeling the co-speech gesture with emotion transition, we first split the existing aligned speech-gesture pairs [26, 33] into four-second clips. Then we randomly splice two clips from the same speaker to construct an emotion-transition candidate pair. The clip with neutral emotion is leveraged as head speech, and the other with various emotions is represented as tail speech. We leverage the dynamically synthesized two-second audio as a transition to combine the head and tail speeches. In the following, we briefly summarize our efforts in constructing the dynamic transition speech audio.

Textual Inpainting of Transition: To ensure the semantic coherence of transition w.r.t. head/tail speeches, we exploit the advanced language generation model ChatGPT-4 [31] to complete the transcript according to the context. Literally, we follow the conventional estimation that people usually talk 30 phonemes [37] in a two-second speech as the prompt to guide transcript generation by ChatGPT-4.

Synthesis of Transition Audio: Once we obtain the transcript of transitions, we employ a superior text-to-speech model, audioLDM2 [27, 28], to generate corresponding speech audio. Here, we leverage the speaker embeddings extracted via SpeechBrain's ECAPA-TDNN [5, 36] as prior guidance to maintain the identity consistency of the generated transition audio. Then, we adopt Whisper [35] to restrict the word error rate, thereby ensuring the accuracy and clarity of the speech content. In this fashion, the synthesized transition audio realizes controlled duration while the natural smooth tonality is well preserved. The datasets will be released in the future. More details are provided in the Section 4.1.

3.2. Problem Formulation

Given a sequence of audio signal $A = \{a_1, ..., a_N\}$ as input to model $\mathcal{G}$, our goal is to generate vivid and emotional 3D co-speech gestures $P = \{p_1, ..., p_N\}$ of the upper human body. N represents the number of synthesized postures corresponding to the audio A. Each p_i is denoted as J joints with 3D representation. In particular, we define the audio signal as consisting of a head speech, a dynamically inpainted transition speech, and a tail speech. Here, the head and tail speeches are randomly selected from previous co-speech gesture datasets [26, 29] with a single emotion label, respectively. Following the conventions of previous works [26, 29, 46, 49], we invoke M frame poses as the initial seed to guide generation. The overall objective is expressed as

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

Note that only the generated gestures with head and tail speeches are supervised with ground truth coming from existing datasets [26, 29]. The transition gestures with L frames, where $L \ll N$, will be weakly supervised through the following processes within a motion transition infusion mechanism and an emotion mixture strategy. The audio signal is fed into an audio encoder for feature extraction. Our overall workflow is shown in Figure 2.

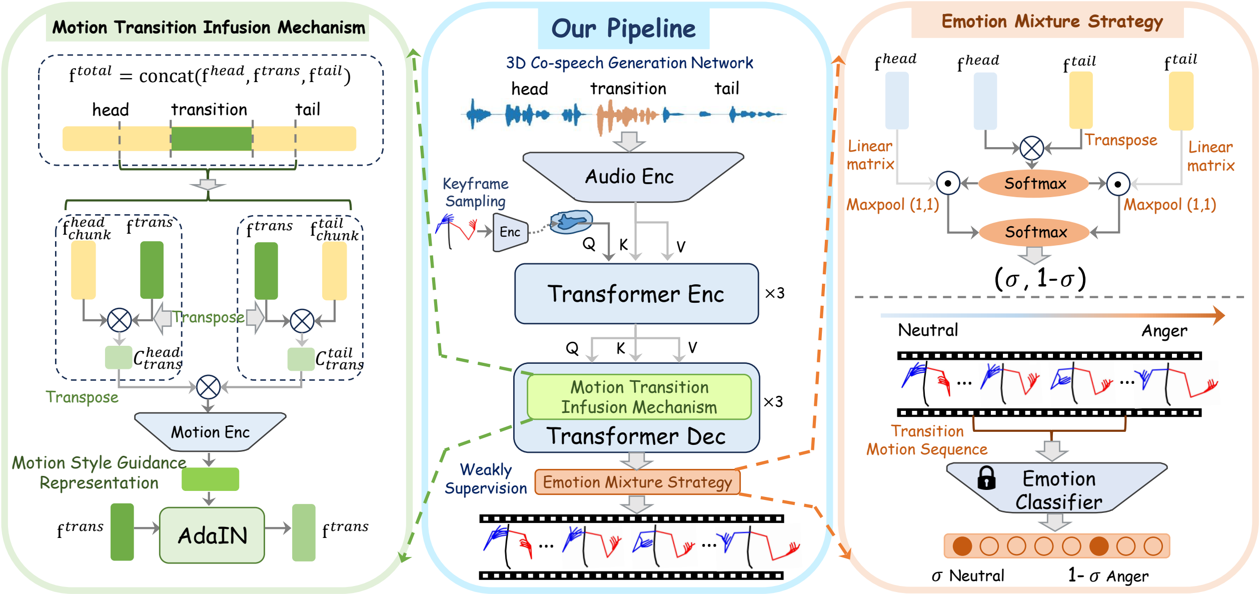

Figure 2. **The overview of our proposed method.** The **middle part** (**blue**) displays the overall pipeline for 3D co-speech gesture generation from emotion transition human speech. The **left part** (**green**) depicts our proposed Motion Transition Infusion Mechanism that enhances the coordination of transition gestures w.r.t. head/tail ones. The **right part** (**orange**) shows the designed Emotion Mixture Strategy to provide weak supervision of the generated transition gestures, thereby achieving authority producing.

3.3. Weakly-supervised Emotion Transition

Motion Transition Infusion Mechanism: To ensure the temporal coherence of the transition gestures w.r.t. head/tail ones, we propose a motion transition infusion mechanism to model the temporal association representation between different emotional gestures. As depicted in Figure 2 (left), the total sequential features $\mathbf{f}^{total}$ consist of the features of head $\mathbf{f}^{head}$, transition $\mathbf{f}^{tran}$, and tail $\mathbf{f}^{tail}$. Inspired by [30], we nominate a head chunk $\mathbf{f}^{head}_{chunk}$ composed of the last L frames of head gesture embeddings and a tail chunk $\mathbf{f}^{tail}_{chunk}$ consisting of the first L ones of the tail gesture embeddings. Here, the dimension of each frame representing gesture embedding is $\mathbb{R}^{1\times D}$.

In particular, we first calculate the temporal correlation matrix $\mathbf{C}^{head}_{trans} \in \mathbb{R}^{L\times L}$ between the head chunk embedding and transition embedding. Here, the temporal correlation matrix represents the temporal variations in the gestures from the head to the transition. Similarly, we obtain the correlation matrix $\mathbf{C}^{tail}_{trans}$ from tail chunk embedding and transition embedding. Then, the global temporal dependency from head to tail is computed via matrix multiplication between $\mathbf{C}^{head}_{trans}$ and $\mathbf{C}^{tail}_{trans}$. Further, we develop a motion encoder to acquire the sequence-aware style guidance representation based on the global temporal dependency. The style guidance representation is exploited to boost the transition embeddings via an adaptive instance normalization (AdaIN) layer [16]. By doing so, we derive the transition $\mathbf{f}^{tran}$ as

$$\mathbf{f}^{tran} = \mathrm{AdaIN}\left\{\mathbf{f}^{tran}, \mathrm{Enc}(\mathbf{C}^{head}_{trans} \otimes \mathbf{C}^{tail}_{trans})\right\}, \quad (2)$$

where $\otimes$ indicates matrix multiplication, and Enc denotes the motion encoder.

Emotion Mixture Strategy: Considering obtaining the realistic 3D pose annotations corresponding to dynamically inpainted transition audio is quite difficult, we design an emotion mixture strategy to provide weak supervision of the generated gestures. Our key insight is built upon the fact that different emotions in the head/tail would lead to different gesture motions, thereby the emotion represented by transition gestures would be a mixture of the head and tail ones. As shown in Figure 2 (right), we utilize two learnable parameters as soft emotion labels of the transition gestures.

Specifically, through computing the correlation matrix between the embeddings of head gestures and tail gestures, we obtain the motion deformation representation S from one emotion to another (*e.g.*, neutral-to-anger). This motion deformation is formulated as:

$$\mathbf{S}^{tail}_{head} = \mathrm{Softmax}(\mathbf{f}^{head} \otimes \mathbf{f}^{tail'}), \quad (3)$$

where $'$ indicates the transpose operation. With the help of the motion deformation representation, we obtain the two learnable emotion embeddings $(\mathbf{e}^{head}, \mathbf{e}^{tail})$ from the head and tail, respectively, as

$$\mathbf{e}^{head} = \mathrm{MaxPool}\left\{\mathbf{W}_{linear}(\mathbf{f}^{head}) \odot \mathbf{S}^{tail}_{head}\right\}, \quad (4)$$

where $\mathbf{W}_{linear}$ denotes linear matrix, and $\odot$ means dot product. Maxpool indicates the AdaptiveMaxPooling operation. $\mathbf{e}^{tail}$ is calculated in a similar way. Then, we obtain the two learnable parameters represented by emotion weights of head and tail gestures, as follows

$$(\sigma, 1-\sigma) = \text{Softmax}(\mathbf{e}^{head}, \mathbf{e}^{tail}). \tag{5}$$

Once we acquire the learnable emotion weights, we leverage a pre-trained pose-based emotion classifier to provide weak supervision. In this fashion, the authority of generated transition gestures is well-preserved. The training details of the pre-trained pose-based emotion classifier are provided in the supplementary material.

Keyframe Posture Sampling: Conventionally, researchers [29, 46, 49] directly leverage the padded initial poses as conditional seeds to guide the co-speech gesture generation. In long-sequence modeling, an intuitive manner is to extend the initial pose length M for adapting the overall synthesized gestures' length. However, this would lead to the poor generalizability of the network (*i.e*., the performance gains degradation from the reduction of initial poses). Therefore, we propose a simple yet effective VAE-based [40] keyframe sampler to provide high-fidelity posture prior conditions while enabling diverse results. Evenly, we split the annotated sequence into several chunks keeping the same length with the transition of length L. The sampler is trained with the keyframe randomly selected to reconstruct the corresponding chunk. In the inference phase, the keyframe sampler samples diverse chunks, blending as the initial postures for gesture generation. Since the transition sequence lacks the pose annotation, we randomly select the frame from head or tail sequences for posture sampling.

To further enhance the sequence-aware correspondence of the generated co-speech gestures, we leverage the diverse initial postures as the query Q to match the key features K and value features V in the transformer-based backbone [41]. Similar to [34], we adopt a motion discriminator to ensure the temporal smoothness of the generated results. For more details about network architecture please refer to supplementary material.

3.4. Objective Functions

Reconstruction Loss: We leverage the ground truth 3D pose annotation of the head and tail to constrain the generated co-speech gestures as:

$$\mathcal{L}_{rec} = \left\| P_{\{head,tail\}} - \widehat{P_{\{head,tail\}}} \right\|_1, \tag{6}$$

where $\widehat{P_{\{head,tail\}}}$ denotes generated gestures of head and tail speeches.

Adversarial Learning Loss: To ensure the realism of the generated gestures, we further exploit the adversarial training loss, expressed as:

$$\mathcal{L}_{adv} = \mathbb{E}_P\left[\log \mathcal{D}(P)\right] + \mathbb{E}_A\left[\log(1 - (\mathcal{G}\left(A, \{p_1, ..., p_M\}\right))\right], \tag{7}$$

where $\mathcal{D}$ denotes the motion discriminator and $\mathcal{G}$ means gesture generator.

Weakly Supervision Loss: We leverage the pre-trained pose-based emotion classifier to provide weak supervision of the transition gestures upon the learnable emotion weights:

$$\mathcal{L}_{emotion} = -y \log \mathcal{F}(P_{trans}), \tag{8}$$

where y is the learnable emotion label, $\mathcal{F}$ is the emotion classifier, $\widehat{P_{trans}}$ is the generated transition gestures. Finally, the overall objective is:

$$\min_G \max_D \mathcal{L}_{total} = \lambda_r \mathcal{L}_{rec} + \lambda_{adv} \mathcal{L}_{adv} + \mathcal{L}_{emotion}. \tag{9}$$

The λ_r, and λ_{adv} are weight coefficients.

4. Experiments

4.1. Datasets and Experimental Setting

BEAT Emotion Transition Dataset (BEAT-ETrans): Since there are only ***single emotion*** labels of aligned speech-gesture corpus in the original BEAT dataset [26], to satisfy our insight on emotion transition co-speech gesture generation modeling, we recollect a BEAT Emotion Transition Dataset (dubbed BEAT-ETrans). In particular, we resample the motion FPS as 15 and intercept the continuous 60 frames with stride 30 as the head/tail clips. Here, the head clips are all annotated as *neutral*, and tail clips are denoted with the other seven emotions: *anger, happiness, fear, disgust, sadness, contempt, and surprise*. As for one head speech, we randomly select two or three tails with different emotions to construct the head-tail pairs. By leveraging the two-second (*i.e*., corresponding 30-frame postures) transition to blend the heads and tails, we obtain the 10-second human speech clips in our BEAT-ETrans. We obtain $58,077$ clips, including a total of 161.3 hours reported in Table 1. Then the clip numbers of training/validation/testing sets are randomly split as $41,908$/$4,077$/$12,092$. In all of our experiments, we utilize the upper body with 71 joints.

TED Emotion Transition Dataset (TED-ETrans): Inspired by [29, 33], we further newly collect a TED Emotion Transition Dataset (dubbed TED-ETrans) based on more than 1.7K speakers from in-the-wild TED talk show videos, demonstrated in Table 1. Due to the lack of emotional labels, we first leverage the annotated BEAT dataset to pretrain an audio-based emotion classifier for labeling TED audios. To ensure the authority of emotion labels, we set the

Table 1. Statistics comparison of existing 3D co-speech gesture datasets with ours. Our **BEAT-ETrans** and **TED-ETrans** are built upon the existing BEAT [26] and TED-Expressive [29], respectively. To the best of our knowledge, we are the first to present two large datasets with emotion transition human speech.

Dataset	Joint Annotation	Modality							Duration (hours)
		Body	Hand	Audio	Text	Speakers	Single Emotion	**Emotion Transition**	
TED [46]	pseudo label	9	✗	✓	✓	1,766	✗	✗	106.1
SCG [13]	pseudo label	14	24	✓	✗	6	✗	✗	33
Trinity [8]	mo-cap	24	38	✓	✓	1	✗	✗	4
ZeroEGGS [10]	mo-cap	27	48	✓	✓	1	✗	✗	2
BEAT [26]	mo-cap	27	48	✓	✓	30	✓	✗	35
TED-Expressive [29]	pseudo label	13	30	✓	✓	1,764	✗	✗	100.8
BEAT-ETrans (ours)	**mo-cap**	**27**	**48**	✓	✓	**30**	**8**	✓	**161.3**
TED-ETrans (ours)	**pseudo label**	**13**	**30**	✓	✓	**1,764**	**6**	✓	**59.8**

Table 2. Comparison with the start-of-the-art methods on our newly collected BEAT-ETrans and TED-ETrans datasets. ↑ denotes the higher the better, and ↓ indicates the lower the better. ± means 95% confidence interval.

Models	BEAT-ETrans				TED-ETrans			
	FGD$_{h+t}$ ↓	FGD$_{trans}$ ↓	BC ↑	Diversity ↑	FGD$_{h+t}$ ↓	FGD$_{trans}$ ↓	BC ↑	Diversity ↑
Seq2Seq [45]$_{ICRA'19}$	40.95	47.93	0.141	$96.66^{\pm 2.16}$	29.60	49.47	0.265	$72.81^{\pm 1.99}$
S2G [11]$_{CVPR'19}$	25.56	37.04	0.671	$98.26^{\pm 2.04}$	18.16	41.63	0.824	$76.82^{\pm 2.32}$
Trimodal [46]$_{TOG'20}$	14.09	42.50	0.764	$100.87^{\pm 2.12}$	21.06	33.20	0.758	$82.87^{\pm 1.86}$
CAMN [26]$_{ECCV'22}$	9.03	27.53	0.794	$118.46^{\pm 2.33}$	19.28	41.04	0.785	$79.03^{\pm 1.49}$
HA2G [29]$_{CVPR'22}$	7.28	25.79	0.779	$121.77^{\pm 2.31}$	16.72	40.38	0.787	$80.14^{\pm 1.65}$
DiffGesture [49]$_{CVPR'23}$	6.68	25.03	0.788	$122.29^{\pm 2.01}$	18.69	25.13	0.818	$92.01^{\pm 2.07}$
Ours	**4.42**	**18.84**	**0.881**	**124.93**$^{\pm 2.10}$	**12.19**	**23.54**	**0.906**	**93.79**$^{\pm 2.53}$

classification threshold as ≥ 0.95, and the two uncommon emotions (*i.e. fear, disgust*) are dropped. Then we maintain the same data pre-processing strategy with our BEAT-ETrans to obtain a total of $21,515$ clips with 59.8 hours. The final clip division criteria of the TED-ETrans dataset are training/validation/testing with 15,061/2,152/4,302 respectively. In practice, the 43 upper body joints are leveraged in the experiments.

Implementation Details: We set the total generated co-speech gesture length as $N = 150$, and the transition and chunk lengths are $L = 30$. Conventionally, we leverage $M = 4$ frames as the reference initial poses. The feature dimension $D = 512$ in practice. The raw audio of human speech is converted to mel-spectrograms with FFT window size 1024, and hop length 512. The audio encoder takes the ResNetSE34 [4] as the backbone. Empirically, we set $\lambda_r = 20$, and $\lambda_{adv} = 2$. Our models are implemented on the Pytorch platform with a single NVIDIA Tesla V100 GPU. The initial learning rate is set to 0.0003 by utilizing Adam Optimizer. The whole training takes 100 epochs with a batch size of 96.

Evaluation Metrics: To fully evaluate the realism and diversity of the generated co-speech gestures, we introduce various metrics*:

- **FGD**: Fréchet Gesture Distance (FGD) [46] is utilized to measure the distribution distance between the realistic sequential gestures and generated ones. Since we only have the 3D joint annotation of the head/tail, we take the network architecture provided in [29, 46] to train the auto-encoder for distance computing on the two datasets, respectively. The FGD of transition gestures is calculated as the average value between the distribution distance of transition and head/tail, indicated as FGD$_{trans}$. Similarly, FGD$_{h+t}$ means the distance between the generated head/tail gestures and ground truth.
- **BC**: Beat Consistency Score (BC) [26, 29] measures the speech audio alignment degree with the generated co-speech gestures.
- **Diversity**: Similar to [29, 49], we exploit the same feature extractor in FGD to obtain the feature embeddings of the generated gestures. The diversity reflects the average distance between 500 random combination pairs in the testing set of 60 speech audios.

4.2. Quantitative Evaluation

Comparisons with SOTA Methods: To the best of our knowledge, we are the first to explore the co-speech gesture generation with emotion transition human audios. To fully verify the superiority of our method, we implement

*More details are in the supplementary material.

Table 3. Ablation study on different components of the proposed method. ✓ indicates the employment of a certain module. ↑ denotes the higher the better, and ↓ indicates the lower the better. ± means 95% confidence interval. MTIM: Motion Transition Infusion Mechanism; EMS: Emotion Mixture Strategy; FKS: Keyframe Sampler.

Model Variations				BEAT-ETrans				TED-ETrans			
Baseline	MTIM	EMS	KFS	FGD$_{h+t}$ ↓	FGD$_{trans}$ ↓	BC ↑	Diversity ↑	FGD$_{h+t}$ ↓	FGD$_{trans}$ ↓	BC ↑	Diversity ↑
✓				21.21	56.33	0.701	$88.89^{\pm 1.58}$	26.44	47.87	0.712	$74.19^{\pm 2.75}$
✓	✓			10.75	27.94	0.866	$108.69^{\pm 2.63}$	20.95	36.38	0.827	$82.42^{\pm 2.24}$
✓	✓	✓		5.69	21.70	0.878	$112.80^{\pm 1.95}$	14.02	26.21	0.900	$87.21^{\pm 2.12}$
✓	✓	✓	✓	**4.42**	**18.84**	**0.881**	**124.93**$^{\pm 2.10}$	**12.19**	**23.54**	**0.906**	**93.79**$^{\pm 2.53}$

various state-of-the-art single-emotion-based counterparts: Seq2Seq [45], S2G [11], Trimodal [46], CAMN [26], HA2G [29], and DiffGesture [49]. For a fair comparison, all the models are implemented by the source code released by the authors. Note that since we only have the ground truth of head and tail gestures, several recent VAE-based works [1, 2, 44, 48] cannot be directly applied in the experiments (or they have not released codes so far).

As reported in Table 2, we adopt the FGD$_{h+t}$, FGD$_{trans}$, BC, and Diversity for a well-rounded view of comparisons. Our method outperforms all the competitors by a large margin on both two datasets. Remarkably, on the TED-ETrans dataset, our method even achieves 34.8% (*i.e.*, $(18.69 - 12.19)/18.69 \approx 34.8\%$) improvement over the sub-optimal counterparts in FGD$_{h+t}$. We observe both the DiffGesture [49] and ours synthesize the high-fidelity gestures of head/tail speech with much lower FGD$_{h+t}$ than others. However, the DiffGesture shows worse performance on FGD$_{trans}$ due to lack of supervision. In terms of diversity, our simple yet effective keyframe sampler provides authority and diverse initial postures as the reference, thus enabling us to demonstrate diverse gesture styles compared to other counterparts. Moreover, we find that the diversity scores on BEAT-ETrans are much higher than those on TED-ETrans dataset. This can be attributed to the more complex human joints in the BEAT-ETrans dataset.

Ablation Study: To further verify the effectiveness of our proposed methods, we conduct the ablation study of different components as variations, reported in Table 3. The baseline model is implemented by a simple transformer-based pipeline with stacking three times blocks in the encoder-decoder. Obviously, all the combinations of our proposed components have positive impacts on the generated results. Specifically, by adding the motion transition infusion mechanism to the baseline, the indicator BC has achieved significant improvement (*e.g.*, 0.701 → 0.866 in the BEAT-Etrans). This result verifies that our motion transition infusion mechanism effectively models the temporal correlation between the transitions w.r.t. head/tail gestures, thus leading to the generated results preserving rhythm coherency with given speech, globally.

Moreover, adopting the emotion mixture strategy ideally improves the performance of FGD$_{trans}$ on both two datasets. This indicates that the learnable emotion mixture wights can provide effective weak supervision by leveraging the pre-trained pose-based emotion classifier. Besides, we have observed significant improvement in the performance of FGD$_{h+t}$ during this phase compared to the previous version. This aligns with our transformer backbone's emphasis on modeling the sequential temporal correlations as a whole. The better transition gestures encourage our model to maintain better temporal consistency, thereby the head/tail gestures achieve better results.

Finally, after additionally employing the keyframe sampler to produce authority initial postures as the reference, our method obtains the best performance. Although the properties of BC and FGD on both datasets just have slightly improvement, the diversity realizes a noticeably better achievement (*e.g.*, 112.80 → 124.93 in the BEAT-ETrans dataset). This highly supports our insight into keyframe-driven diversification strategy.

4.3. Qualitative Evaluation

Visualization: To fully demonstrate the performance of our method, we show the visualized keyframes generated from ours compared with counterparts on our newly collected TED-ETrans and BEAT-ETrans datasets, respectively. As depicted in Figure 3, our method displays vivid and diverse results against others. In particular, we observe that Seq2Seq and Trimodal tend to synthesize unreasonable and stiff results (*e.g.*, the blue rectangle in the right BEAT-ETrans dataset). Although CAMN and HA2G can generate natural upper-body postures, we find that they sometimes produce unreliable subtle fingers (*e.g.*, the red rectangle in the left TED-ETrans dataset). Both our method and DiffGesture create reasonable gestures. However, the results synthesized by DiffuGesture are mismatched with the emotion transitions. In contrast, our method can synthesize the synchronous motions (*e.g.*, in the BEAT-ETrans, the arms become droopy as emotion turns to sadness). Meanwhile, we further verify the diversification results as shown in Figure 1. Given the same input audio, our method generates diverse and vivid co-speech gestures. Please refer to the

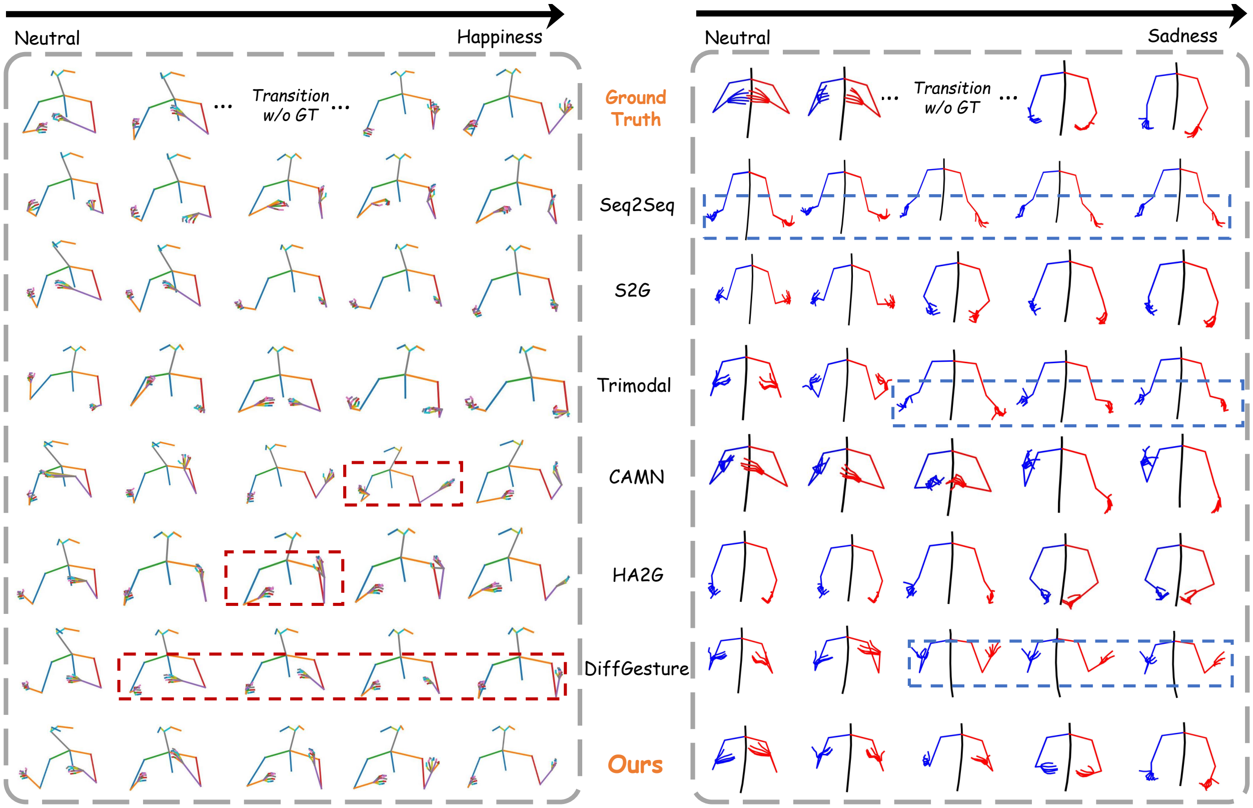

Figure 3. Visualization of our generated 3D co-speech gestures against various state-of-the-art methods. The samples of the left part are from our newly collected TED-ETrans dataset, and the samples of the right part are from our BEAT-ETrans dataset. Best view on screen.

supplementary material for more visualization results.

User Study: To further analyze the quality of results by various counterparts and ours, we conduct a user study by visualizing the results and inviting 15 subjects. In particular, the subjects are required to rate the generated co-speech gestures from 0 to 5 (the higher, the better) in terms of naturalness, motion smoothness, and speech-gesture coherency. The results are demonstrated in Figure 4. Our method showcases the best performance compared with all the competitors. Especially in terms of motion smoothness, our method achieves noticeable advantages, indicating the effectiveness of our proposed motion transition infusion mechanism and the emotion mixture strategy.

5. Conclusion

In this paper, we introduce a new task of 3D co-speech gesture generation given emotion transition human speech. We therefore newly collected two datasets named the BEAT-ETrans and the TED-ETrans to fulfill this goal while significantly facilitating the research on 3D human motion modeling. Then, we fully take advantage of the sequential temporal correlation via a motion transition infusion mechanism to ensure the generated gestures preserve temporal coherence. Furthermore, we design an emotion mixture strategy to supply emotional weak supervision of the synthesized transitions. Extensive experiments conducted on our two newly collected datasets show the superiority of the method. As our method intends to generate diverse and vivid emotion transition gestures, we will investigate diversifying the 3D gesture with temporal smooth sampling, instead of the keyframe-wise manner.

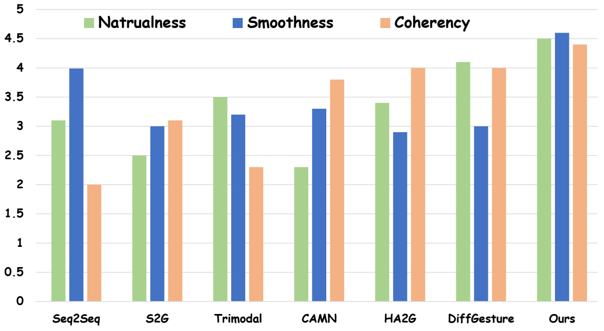

Figure 4. User study on gesture naturalness, motion smoothness, and speech-gesture coherency.

Weakly-Supervised Emotion Transition Learning for Diverse 3D Co-speech Gesture Generation

Supplementary Material

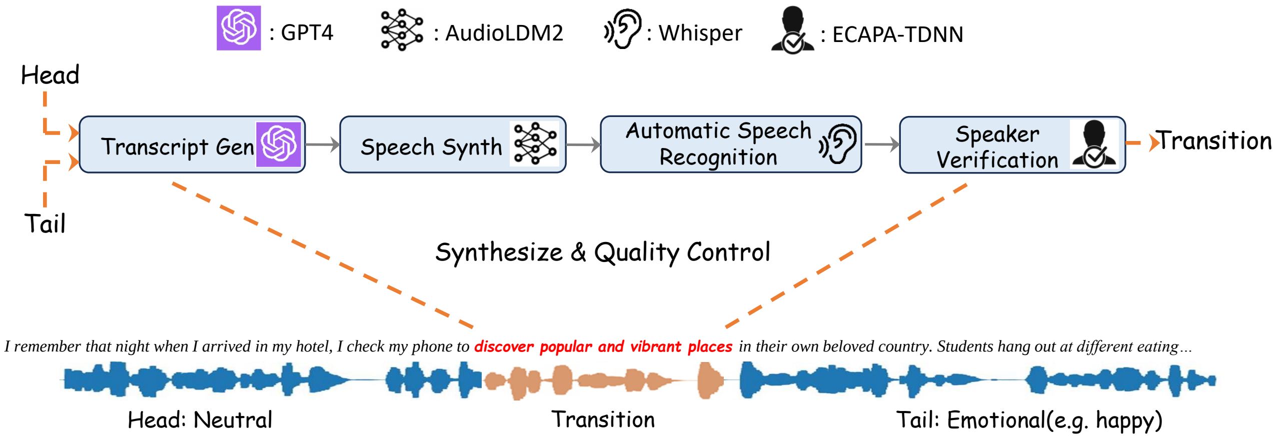
: GPT4
: AudioLDM2
: Whisper
: ECAPA-TDNN
Head
Tail
Transcript Gen
Speech Synth
Automatic Speech Recognition
Speaker Verification
Transition
Synthesize & Quality Control
I remember that night when I arrived in my hotel, I check my phone to discover popular and vibrant places in their own beloved country. Students hang out at different eating…
Head: Neutral
Transition
Tail: Emotional(e.g. happy)

Figure 5. The pipeline of dataset construction. Head and tail audios as well as the corresponding transcripts are fed into the pipeline to generate a smooth and high quality transition.

6. Overview

To demonstrate the effectiveness of our data construction techniques and the proposed method of emotion transition co-speech gesture generation, we further elaborate on the detailed data synthesis and vision perception in the supplementary material. The additional content is illustrated in the following folds:

- Dataset Construction
- Architecture Details
- Additional Experiments

7. Dataset Construction

We will release our newly collected the TED-ETrans and BEAT-ETrans datasets in the future. The overall pipeline of our approach to constructing the dataset is displayed in Figure 5. The details involve the following steps:

Segmentation and Emotion Labeling : We first divide the previously aligned single emotion co-speech gesture datasets [26, 29] into head and tail segments by splitting the original audio into 4-second clips. Heads are identified as clips with neutral emotions, while tails contain various emotions. This segmentation was achieved using either the pre-annotated dataset's emotion labels or an emotion classifier. Both head and tail segments originated from the same speaker, ensuring vocal tone consistency.

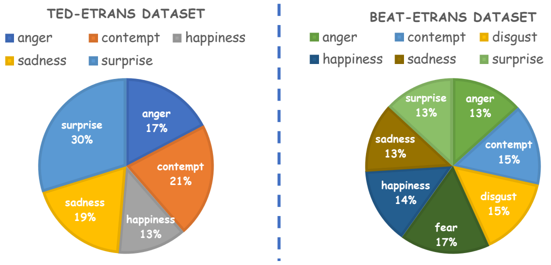

Figure 6. Details of emotion transition distribution of our newly collected TED-ETrans and BEAT-ETrans datasets. All the transitions start from the neutral emotional speeches.

Emotion Transition : The head segments consistently exhibit neutral emotions, while the tails display a variety of emotional states. In our approach, we intentionally avoided pairing segments with extreme emotional shifts (*e.g.*, happiness-to-anger, happiness-to-sadness). Such drastic transitions are infrequent in natural speech and not only result in less smooth transitions in both speech and textual contexts but also risk introducing a long-tail phenomenon in the dataset. By avoiding these extremes, we aimed to maintain a more balanced and realistic dataset distribution as shown in Figure 6.

Transcript Generation with GPT-4 : We engage GPT-4 to generate transitional text between the head and tail clips. The GPT-4 is instructed to create a smooth transition in both content and emotion, producing about 5-10 words. For each data sample, GPT-4 generated three candidate transitions, each accompanied by a confidence score, returned in JSON format. We finally discard samples with low confidence or excessive length.

Synthesis of Transition Speech : We employ the AudioLDM2 [27, 28] model for audio inpainting, ensuring natural and time-controlled speech synthesis. Speaker embeddings are extracted using SpeechBrain's ECAPA-TDNN to measure the consistency of the transition speech with the head and tail segments. Samples with significant speaker embedding discrepancies are excluded. We ensure the head, tail, and synthesized parts share the same speaker's tone, maintaining consistency.

Quality Control through ASR : We utilize Whisper [35] for automatic speech recognition (ASR) on transition speech. ASR transcripts are compared to ground truth, and samples with the word error rate of over 0.125 are re-synthesized for better accuracy and clarity.

Final Note : We observe that GPT-3.5 often produces similar candidates, lacking diversity, thus our preference for GPT-4. Our final prompt structure, designed to guide the model in generating contextually and emotionally coherent transitions, is presented below:

Prompt

As a skilled playwright, you've been assigned a task to fill in the blanks. You will be given two sentences (in a talk) with distinct emotions, and your job is to provide a transition of ∗∗10 words∗∗ to ensure a natural emotional and semantic flow between them. For each blank, you should return three potential options along with your confidence level in your responses in JSON format. ∗∗DO NOT∗∗ return anything else.
JSON template:
{
”opt1”: option 1,
”opt2”: option 2,
”opt3”: option 3,
”confi”: confidence score scale from 1 to 5,
}

Input:\n

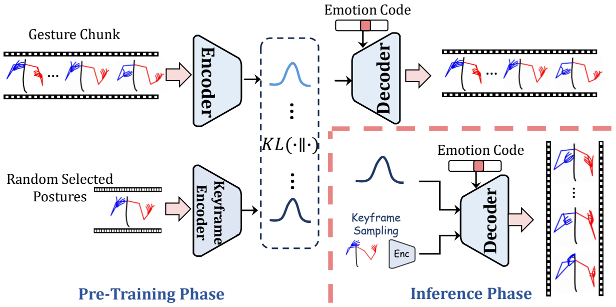

Figure 7. Details of our proposed keyframe sampling strategy. Once we obtain the pre-trained keyframe encoder, we leverage it to model the conditional distribution, producing diverse initial postures as the reference.

8. Architecture Details

Audio Encoder. Inspired by [4, 29, 49], the backbone of our audio encoder E_a is constructed as ResNetSE34. 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.

Transformer-based backbone. We leverage the diversified authority initial postures to interact with the extracted audio features. In particular, we leverage the pose reference Q to match the key features K and value features V in the transformer-based encoder via three times Multi-Head Attention (MHA) [41], expressed as:

$$MultiHead(Q, K, V) = softmax(\frac{QK}{\sqrt{d}})V, \tag{10}$$

where d is a normalization constant.

Pose-based Emotion Classifier. In our emotion mixture strategy, we pre-train a pose-based emotion classifier for providing emotional weak supervision on the generated transition gestures. Specifically, the emotion classifier directly leverages the transformer backbone, the same as the pipeline encoder, to extract the sequential pose features. Then, we utilize an MLP-based classifier header on the pose gestures to produce the final emotion categories. In the BEAT-ETrans dataset, our pre-trained emotion classifier achieves 99.92% accuracy. In the TED-ETrans dataset, the accuracy is 99.26%.

Keyframe Sampler. We design a simple but effective VAE-based keyframe sampler to produce authority initial postures as motion cues, thereby facilitating the diversification of the generated 3D co-speech gestures. As shown in Figure 7, the keyframe sampler aims to model the conditional distribution upon the given randomly selected postures. In the pre-training phase, the posterior distribution is denoted as the latent variable from the encoded chunk-wise gestures. The prior distribution of this latent variable is modeled by the keyframe encoder. The training goal in this phase is to minimize the distance between the posterior distribution and the prior one via KL divergence represented as $KL(\cdot \parallel \cdot)$ in Figure 7. Meanwhile, we exploit the L_1 loss to constrain the reconstructed chunk-wise gestures.

9. Additional Experiments

9.1. Metric Calculation Details

Inspired by [29, 46], we take FGD to evaluate whether the generated gestures maintain realism with the ground truth ones in the perceptive of distribution. Conventionally, the feature extractor of FGD is calculated to embed overall sequential gestures into latent space and then utilize a decoder for reconstruction. However, since we do not have the ground truth of the transition gestures, we newly pre-train the feature extractor with the transition length L. In the inference stage, FGD$_{h+t}$ is calculated by averaging the distances between five randomly selected chunks of length L from the head/tail and GT, respectively. Similarly, FGD$_{trans}$ is computed as the average value between the distance of transition and five randomly selected chunks of head/tail. ***We will release the code of our pipeline and evaluation metrics in the future.***

9.2. Additional Visualization Results

Here, we provide more visual results of our methods compared with other counterparts in the ***demo video.*** Meanwhile, to fully demonstrate the effectiveness of our proposed components in the ablation study, we visualize vital frames of the synthesized gestures. As illustrated in Figure 8 and Figure 9, we can clearly observe that all the combinations of our proposed components have positive impacts on the generated results.

References

[1] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics (TOG)*, 41(6):1–19, 2022. 7

[2] Tenglong Ao, Zeyi Zhang, and Libin Liu. Gesturediffuclip: Gesture diffusion model with clip latents. *arXiv preprint arXiv:2303.14613*, 2023. 2, 3, 7

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

[4] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han. In defence of metric learning for speaker recognition. *arXiv preprint arXiv:2003.11982*, 2020. 6, 2

[5] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*, 2020. 3

[6] Mark Endo, Joy Hsu, Jiaman Li, and Jiajun Wu. Motion question answering via modular motion programs. *arXiv preprint arXiv:2305.08953*, 2023. 3

[7] Maged Farouk. Studying human robot interaction and its characteristics. *International Journal of Computations, Information and Manufacturing (IJCIM)*, 2(1), 2022. 2

[8] Ylva Ferstl and Rachel McDonnell. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 93–98, 2018. 6

[9] Yu Fu, Yan Hu, and Veronica Sundstedt. A systematic literature review of virtual, augmented, and mixed reality game applications in healthcare. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(2):1–27, 2022. 2

[10] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F Troje, and Marc-André Carbonneau. Zeroeggs: Zero-shot example-based gesture generation from speech. In *Computer Graphics Forum*, pages 206–216. Wiley Online Library, 2023. 6

[11] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019. 6, 7

[12] Susan Goldin-Meadow and Martha Wagner Alibali. Gesture's role in speaking, learning, and creating language. *Annual review of psychology*, 64:257–283, 2013. 1

[13] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 101–108, 2021. 6

[14] Autumn B Hostetter and Martha W Alibali. Visible embodiment: Gestures as simulated action. *Psychonomic bulletin & review*, 15:495–514, 2008. 1

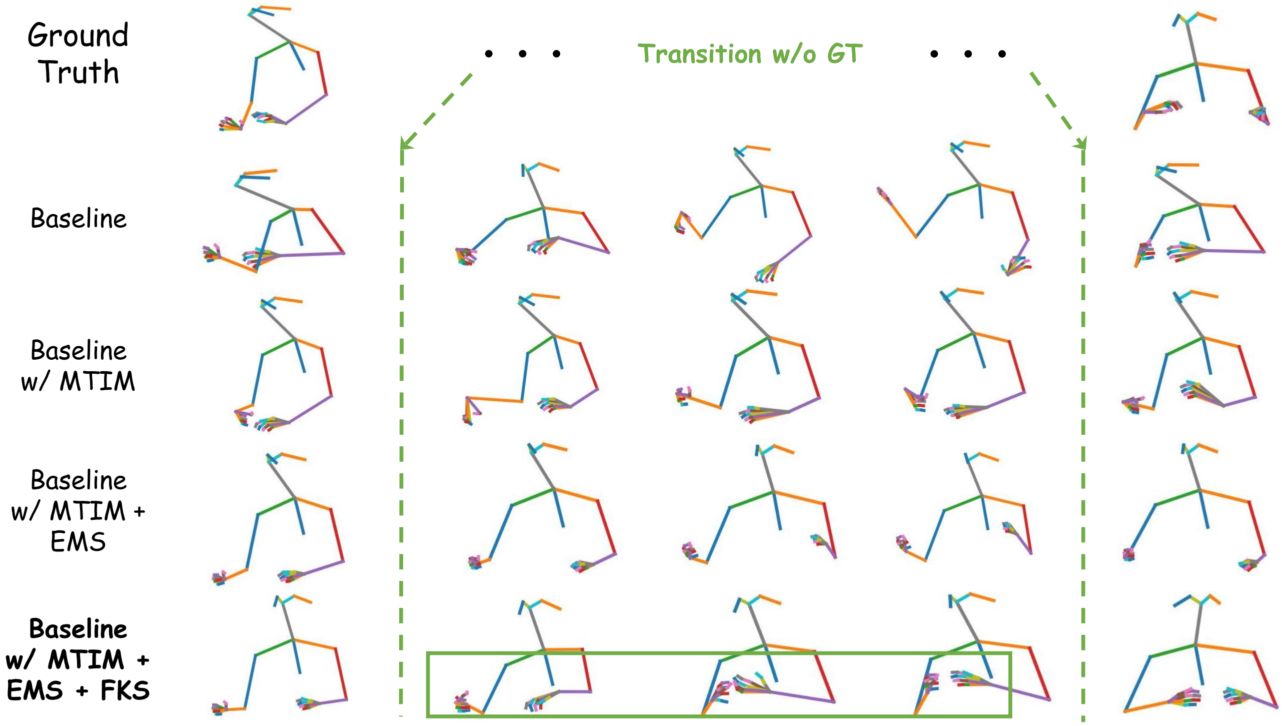

Figure 8. Visual comparisons of ablation study on our newly collected **TED-ETrans dataset**. We show the key frames of the generated motions given the emotion transition of human speech. Best view on screen.

[15] Ann Huang, Pascal Knierim, Francesco Chiossi, Lewis L Chuang, and Robin Welsch. Proxemics for human-agent interaction in augmented reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2022. 2

[16] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 2, 4

[17] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *arXiv preprint arXiv:2306.14795*, 2023. 3

[18] Michael Kipp. *Gesture generation by imitation: From human behavior to computer character animation*. Universal-Publishers, 2005. 2

[19] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. Towards a common framework for multimodal generation: The behavior markup language. In *Intelligent Virtual Agents: 6th International Conference, IVA 2006, Marina Del Rey, CA, USA, August 21-23, 2006. Proceedings 6*, pages 205–217. Springer, 2006. 2

[20] Hema Swetha Koppula and Ashutosh Saxena. Anticipating human activities for reactive robotic response. In *IROS*, page 2071. Tokyo, 2013. 2

[21] Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. Gesture controllers. In *Acm siggraph 2010 papers*, pages 1–11. 2010. 2

[22] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. Learning to generate diverse dance motions with transformer. *arXiv preprint arXiv:2008.08171*, 2020. 3

[23] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11293–11302, 2021. 3

[24] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 3

[25] Chen Liu, Peike Patrick Li, Xingqun Qi, Hu Zhang, Lincheng Li, Dadong Wang, and Xin Yu. Audio-visual segmentation by exploring cross-modal mutual semantics. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7590–7598, 2023. 2

[26] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European Confer-*

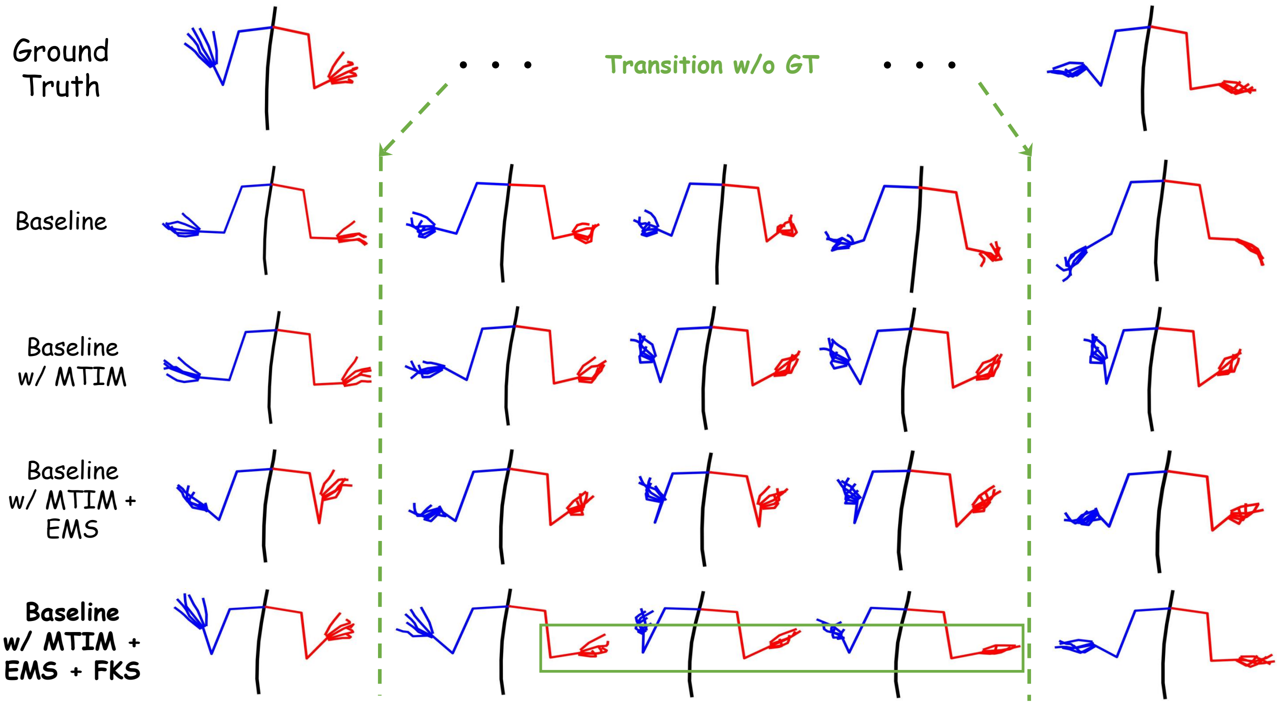

Figure 9. Visual comparisons of ablation study on our newly collected **BEAT-ETrans dataset**. We show the key frames of the generated motions given the emotion transition of human speech. Best view on screen.

ence on Computer Vision, pages 612–630. Springer, 2022. 2, 3, 5, 6, 7, 1

[27] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. *Proceedings of the International Conference on Machine Learning*, 2023. 2, 3

[28] Haohe Liu, Qiao Tian, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *arXiv preprint arXiv:2308.05734*, 2023. 2, 3

[29] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10462–10472, 2022. 2, 3, 5, 6, 7, 1

[30] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Weakly-supervised action transition learning for stochastic human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8151–8160, 2022. 3, 4

[31] OpenAI. Gpt-4 technical report, 2023. 2, 3

[32] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021. 3

[33] Xingqun Qi, Chen Liu, Lincheng Li, Jie Hou, Haoran Xin, and Xin Yu. Emotiongesture: Audio-driven diverse emotional co-speech 3d gesture generation. *arXiv preprint arXiv:2305.18891*, 2023. 2, 3, 5

[34] Xingqun Qi, Chen Liu, Muyi Sun, Lincheng Li, Changjie Fan, and Xin 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*, pages 4616–4626, 2023. 2, 5

[35] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023. 3, 2

[36] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*, 2021. 3

[37] Peter Roach. *Some languages are spoken more quickly than others*. 1998. 3

[38] Mehmet E Sargin, Yucel Yemez, Engin Erzin, and Ahmet M Tekalp. Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1330–1345, 2008. 2

[39] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando:

3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11050–11059, 2022. 3

[40] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015. 5

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5, 2

[42] Jordan Voas, Yili Wang, Qixing Huang, and Raymond Mooney. What is the best automated metric for text to motion generation? *arXiv preprint arXiv:2309.10248*, 2023. 3

[43] Zijian Wang, Xingqun Qi, Kun Yuan, and Muyi Sun. Self-supervised correlation mining network for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7703–7712, 2022. 2

[44] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 469–480, 2023. 2, 3, 7

[45] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4303–4309. IEEE, 2019. 6, 7

[46] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39 (6):1–16, 2020. 2, 3, 5, 6, 7

[47] Weiyu Zhao, Liangxiao Hu, and Shengping Zhang. Diffugesture: Generating human gesture from two-person dialogue with diffusion models. In *International Cconference on Multimodal Interaction*, pages 179–185. 2023. 3

[48] Yihao Zhi, Xiaodong Cun, Xuelin Chen, Xi Shen, Wen Guo, Shaoli Huang, and Shenghua Gao. Livelyspeaker: Towards semantic-aware co-speech gesture generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20807–20817, 2023. 2, 7

[49] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10544–10553, 2023. 2, 3, 5, 6, 7