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

Supplementary Material

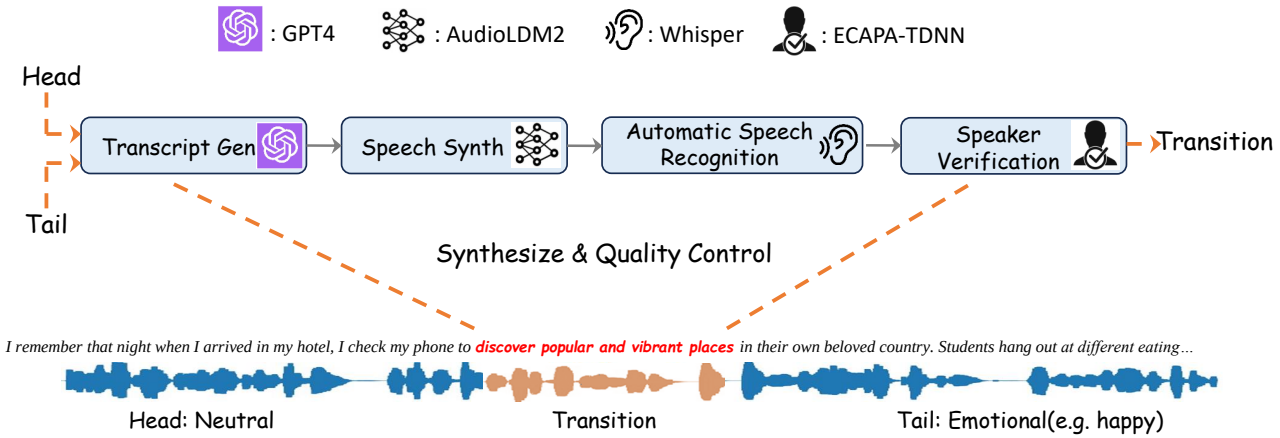

Figure 1. The pipeline of dataset construction. Head and tail audios as well as the corresponding transcripts are fed into the pipeline to generate a smooth and high quality transition.

1. 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

2. 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 1. The details involve the following steps:

Segmentation and Emotion Labeling : We first divide the previously aligned single emotion co-speech gesture datasets [2, 5] 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.

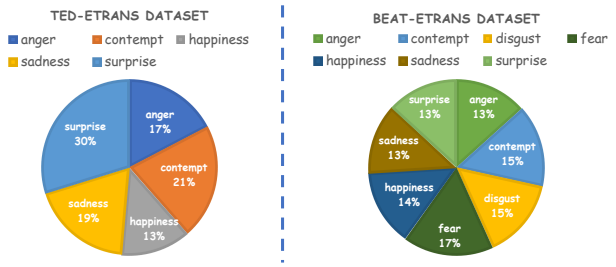

Figure 2. Details of emotion transition distribution of our newly collected TED-ETrans and BEAT-ETrans datasets. All the transitions start from the neutral emotional speeches.

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 2.

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 [3, 4] 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 [6] 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

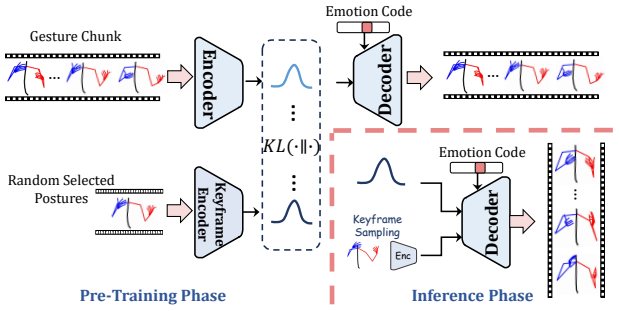

Figure 3. Details of our proposed keyframe sampling strategy. Once we obtain the pre-trained keyframe encoder, we leverage it to model the conditional distribution, producing diverse initial postures as the reference.

We first define GPT-4 prompt and evaluate sentence completeness **confidence scores** three times to select the best fit for semantic clarity. Moreover, we add a manual review on each transcript to drop the unnatural sentences, including 30 English native speakers' evaluation of grammatical correctness/ logical coherence/ clarity of expression. We are unable to extract the transitional segments directly from lengthy videos due to the absence of speech recognizers for identifying multiple emotions within a single audio. Thus, we leverage the advanced LLM GPT-4 combined with the manual effort to construct the natural and smooth emotion transition datasets.

3. Architecture Details

Audio Encoder. Inspired by [1, 5, 9], 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) [7], expressed as:

$$MultiHead(Q, K, V) = softmax(\frac{QK}{\sqrt{d}})V, \quad (1)$$

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 3, 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 3. Meanwhile, we exploit the L_1 loss to constrain the reconstructed chunk-wise gestures.

4. Additional Experiments

4.1. Metric Calculation Details

Inspired by [5, 8], 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.***

4.2. Additional Ablation Study Experiments

As reported in Table 1, after adding the adversarial loss, FGD_{trans} and BC achieve better results. This highly aligns with our motivation to ensure the temporal smoothness of the generated results. Inspired by BEAT, we leverage a pre-trained posture-based emotion classifier to evaluate the emotion transition effect in both datasets. As reported in Table 2, our method attains the best performance on emotion transition, which highly aligns with our visualization.

Table 1. Ablation study on the adversarial loss in TED-ETrans dataset. w/o represents without adversarial loss in experiments.

Methods	$FGD_{h+t} \downarrow$	$FGD_{trans} \downarrow$	BC ↑	Diversity ↑
Ours w/o	15.31	32.72	0.802	$79.64^{\pm 4.58}$
Ours	**12.19**	**23.54**	**0.906**	$\mathbf{93.79}^{\pm 2.53}$

Table 2. Comparison in emotion transition effect. EmoACC means whether the gestures in the head/tail represent the corresponding emotions.

Models	BEAT-ETrans				TED-ETrans			
	FGD$_{h+t}$↓	FGD$_{trans}$↓	BC↑	EmoACC↑	FGD$_{h+t}$↓	FGD$_{trans}$↓	BC↑	EmoACC↑
Seq2Seq	40.95	47.93	0.141	57.50	29.60	49.47	0.265	56.20
S2G	25.56	37.04	0.671	60.69	18.16	41.63	0.824	58.88
Trimodal	14.09	42.50	0.764	69.81	21.06	33.20	0.758	63.82
CAMN	9.03	27.53	0.794	72.87	19.28	41.04	0.785	74.55
HA2G	7.28	25.79	0.779	73.98	16.72	40.38	0.787	80.74
DiffGesture	6.68	25.03	0.788	80.72	18.69	25.13	0.818	81.17
Ours	**4.42**	**18.84**	**0.881**	**83.57**	**12.19**	**23.54**	**0.906**	**85.61**

4.3. 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 4 and Figure 5, we can clearly observe that all the combinations of our proposed components have positive impacts on the generated results.

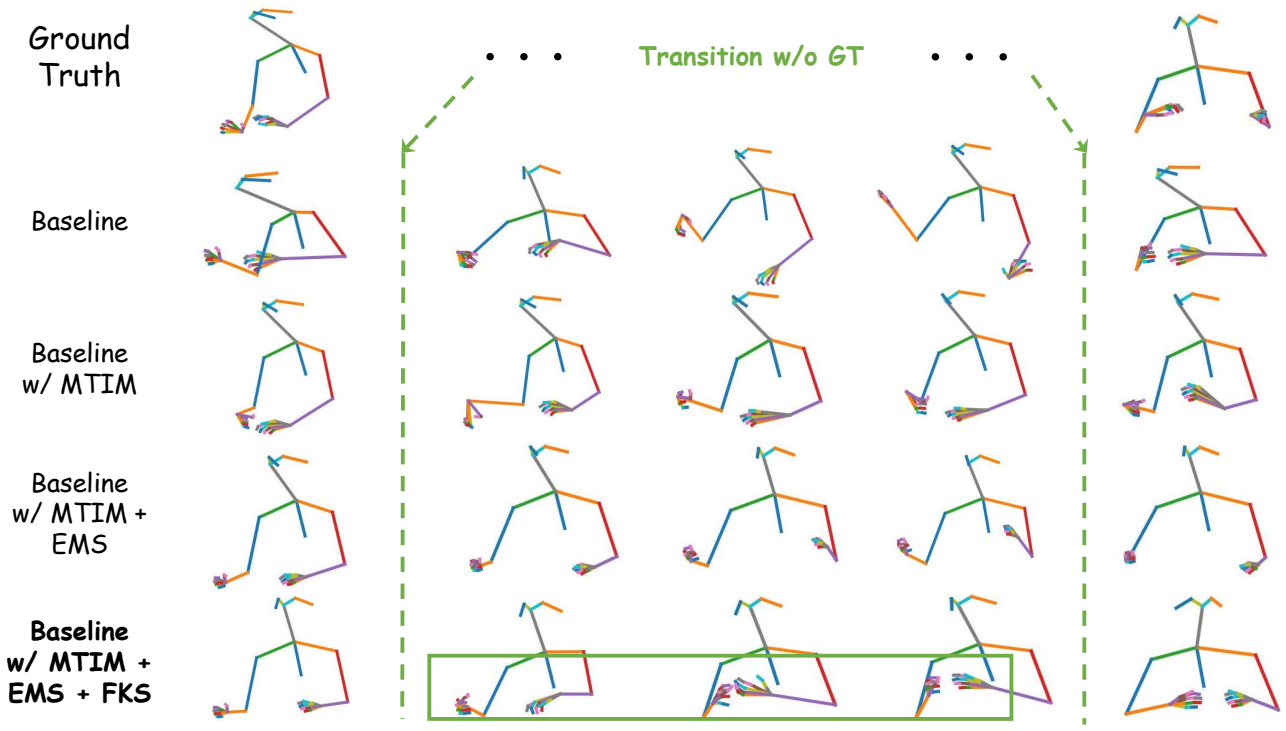

Figure 4. Visual comparisons of ablation study on our newly collected **TED-ETrans dataset**. We show the key frames of the generated motions given the emotion transition of human speech. Best view on screen.

References

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

[2] 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 Conference on Computer Vision*, pages 612–630. Springer, 2022. 1

[3] 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

[4] 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

[5] 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. 1, 2, 3

[6] 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. 2

[7] 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. 2

[8] 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. 3

[9] 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

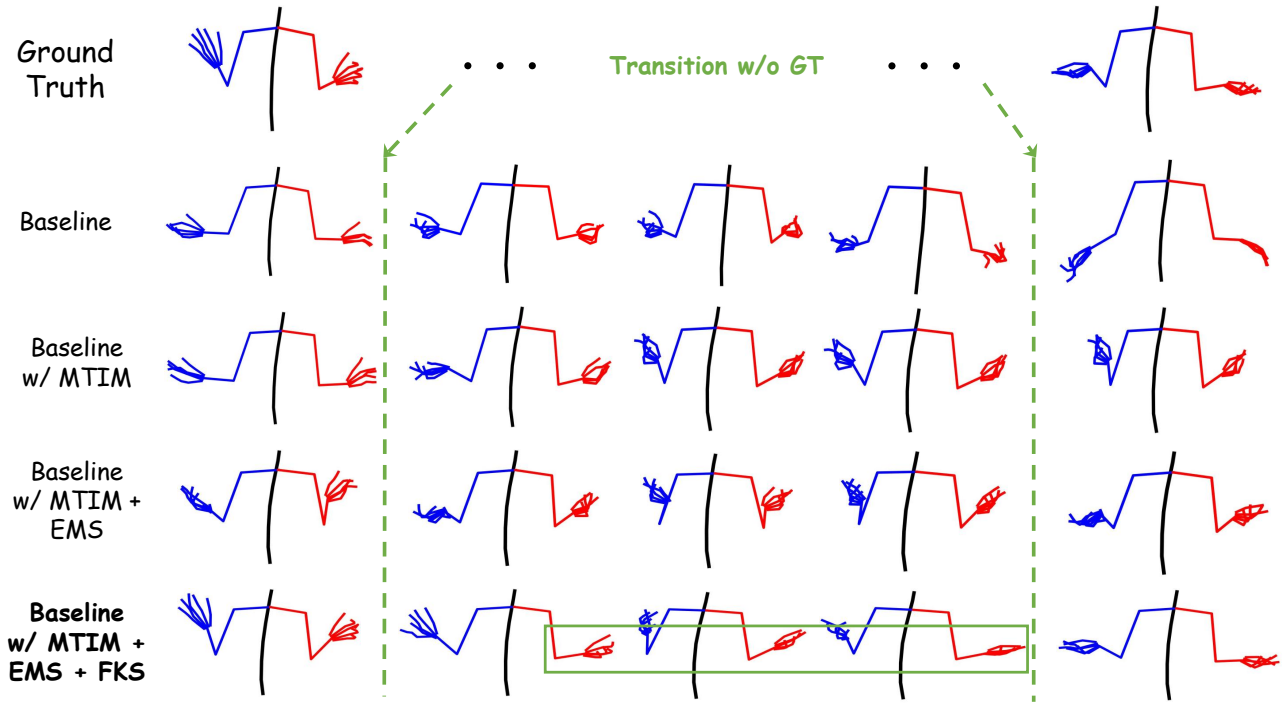

Figure 5. Visual comparisons of ablation study on our newly collected **BEAT-ETrans dataset**. We show the key frames of the generated motions given the emotion transition of human speech. Best view on screen.