RECOM: REALISTIC CO-SPEECH MOTION GENERATION WITH RECURRENT EMBEDDED TRANSFORMER

Yong Xie[1*], *Yunlian Sun*[1*†], *Hongwen Zhang*[2], *Yebin Liu*[3], *Jinhui Tang*[4]

[1]Nanjing University of Science and Technology, [2]Beijing Normal University, [3]Tsinghua University,
[4]Nanjing Forestry University

ABSTRACT

We present ReCoM, an efficient framework for generating high-fidelity and generalizable human body motions synchronized with speech. The core innovation lies in the Recurrent Embedded Transformer (RET), which integrates Dynamic Embedding Regularization (DER) into a Vision Transformer (ViT) core architecture to explicitly model co-speech motion dynamics, enabling joint spatial-temporal dependency modeling. We enhance the model's robustness, noise resistance and cross-domain generalization. To mitigate inherent limitations of autoregressive inference, including error accumulation and limited self-correction, we propose an iterative reconstruction inference (IRI) strategy, which refines motion sequences via cyclic pose reconstruction. Extensive experiments on benchmark datasets validate ReCoM's effectiveness, achieving state-of-the-art performance across metrics. Notably, it reduces the Fréchet Gesture Distance (FGD) from 18.70 to 2.48, demonstrating an 86.7% improvement in motion realism.

Index Terms— co-speech gesture generation, VQ-VAE, vision transformer, zero-shot generation, human motion generation

1. INTRODUCTION

Human motion generation is a broad concept aimed at creating natural and realistic human motions, including whole-body movements such as walking, dancing, etc. Recently, many excellent works have emerged, such as [1, 2, 3, 4, 5, 6], all of which have made outstanding contributions to the field of human motion generation.

Co-speech gesture generation is an important subtask in human motion generation. It involves the automatic generation of expressions and gestures that are seamlessly aligned with speech, and it primarily focuses on upper-body movements, including the motion of the body, hands, and face. This research area holds significant importance for human-computer interaction, digital entertainment, virtual reality, intelligent robots, and other domains. These gestures can help to enrich speech presentation, thus achieving a more natural and fascinating communication experience.

Presently, the field has drawn significant attention. Studies such as those in [7, 8] use VQ-VAE. Studies in [9, 8] manage to perform the step-by-step processing and transformation of data by means of cascaded architectures. The study in [10] takes DiT [11] as its main architecture. Furthermore, studies [12, 13, 14] adopt diffusion models as their principal architectures. Moreover, the research in [15] introduces product quantization to the VAE and enriches the representation of complex holistic motion.

Though achieving impressive results, these approaches face challenges. One is ensuring that the model learns appropriate gesture features meeting the demands of fidelity and generalization. Another is that large gesture datasets are hard to acquire. Fortunately, many good pose estimation methods were proposed [16, 17, 18], fulfilling the need for acquiring high-quality datasets. Hence, we emphasize addressing the first challenge.

Our motivation for ReCoM mainly stems from two observations. Firstly, we noticed that previous methods don't adequately fit the dataset, meaning these models might lack sufficient learning ability. For instance, we observed that the Habibie et al.[19] method generates gestures with excessively large jitter amplitudes. TalkSHOW [7], which occasionally produces jittery motions, frozen movements, and noticeable penetration in animation, results in low fidelity. Secondly, we found that prior methods lack robust generalization capabilities, as evidenced by their poor performance on out-of-domain datasets. For instance, ProbTalk [15] demonstrates suboptimal performance on generalization datasets, and sometimes its visualized motions are overly slow. Given humans' strong perceptual sensitivity to unnatural human motions, these issues may lead to a poor user experience when applied.

Due to the aforementioned issues, we propose corresponding improvement strategies. Our model's competitive edge, as demonstrated in Section 4, stems from three key improvements. (1) We innovatively design the RET module atop the ViT[20] architecture, as in Figure 1, enabling joint spatial-temporal dependency modeling. (2) We propose dynamic Embedding Regularization (DER), a key data processing strategy. DER applies dropout [21, 22] after the embedding layer, active during training but inactive during inference. This reduces complex co-adaptations in large models [21], introduces noise to enhance robustness, and improves generalization. (3) We introduce an Iterative Reconstruction Inference (IRI) strategy for the inference phase, specifically designed to address the inherent limitations of autoregressive inference.

In summary, our contributions are as follows:

- Our work leverages the structural characteristics of ViT to design RET, which enables the model to effectively perceive and process spatio-temporal information. Additionally, we preserve the scalability of ViT and its compatibility with other models, providing insights for adopting appropriate model structures for gesture generation.
- By deploying effective strategies, we significantly enhance the model's learning capability and effectively improve its generalization ability, thereby preventing the model from generating frozen movements or penetration artifacts in animations when handling out-of-domain audio inputs.
- To address the limitations of autoregressive inference, we introduce a novel inference strategy (IRI), prudently adopt the

* indicates equal contribution. †denotes corresponding author (email: yunlian.sun@njust.edu.cn). Thanks to the National Natural Science Foundation of China under Grants 62476131 and 62377004 for funding.

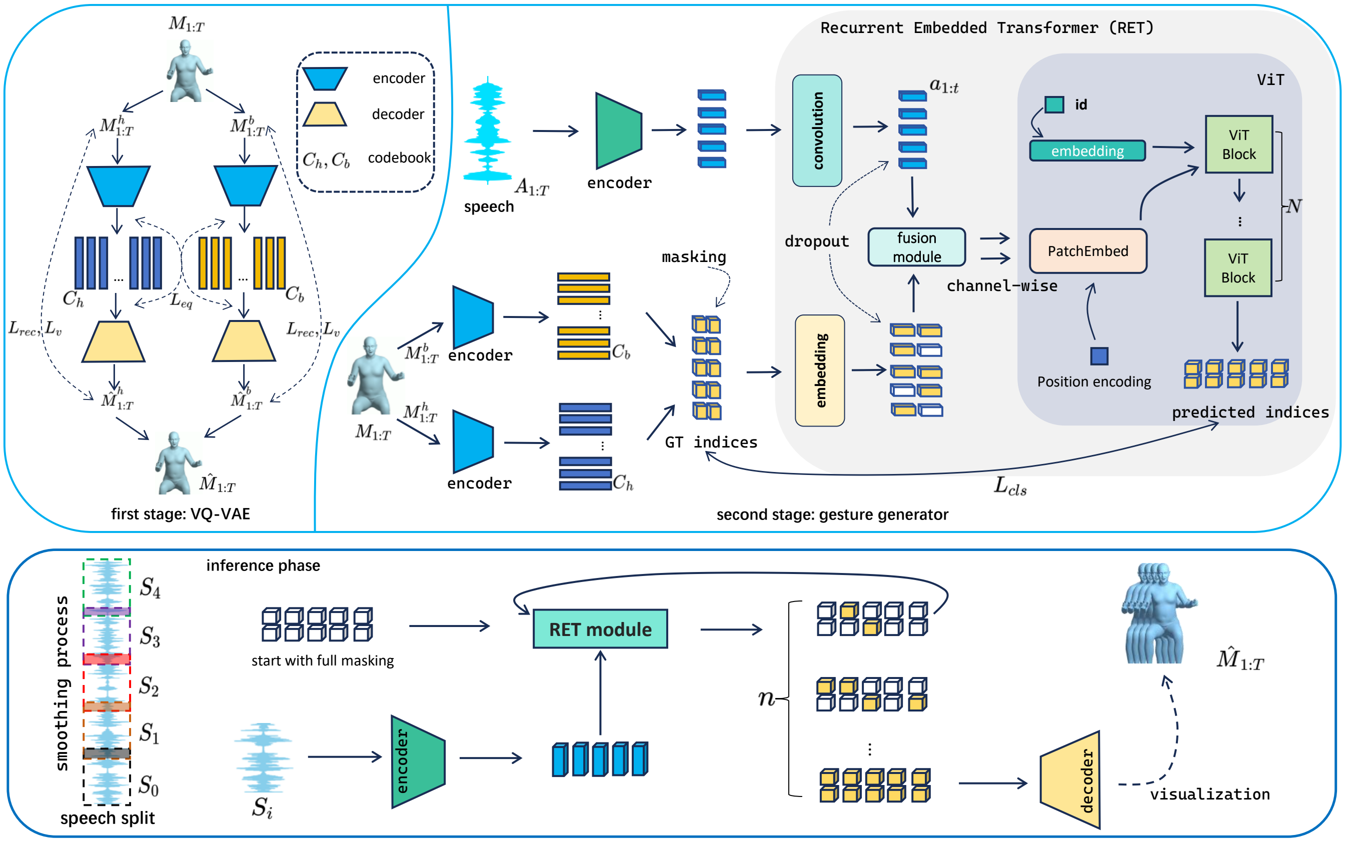

Fig. 1. Training and inference of ReCoM.

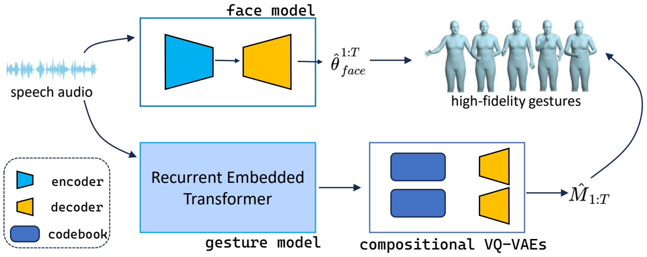

Fig. 2. Our ReCoM pipeline.

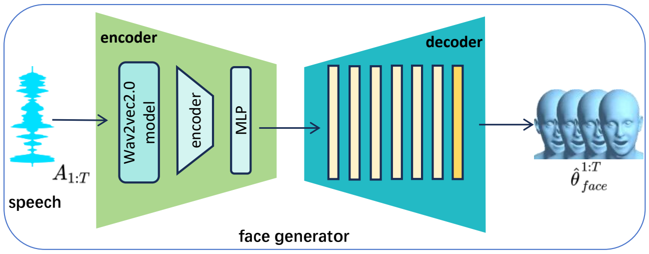

Fig. 3. Face generator

CFG strategy, and explore a temporal smoothing process tailored to Transformer-encoder models. These strategies have a positive impact on the model.

2. METHOD

2.1. Pipeline Overview

Given a speech recording, our ReCoM aims to generate high-fidelity gestures corresponding to it. The overall pipeline is in Figure 2. In our method, we use $\theta_{face} \in \mathbb{R}^{103}$ to represent face parameters. It consists of θ_j and θ_e, where $\theta_j \in \mathbb{R}^3$ represents jaw pose and $\theta_e \in \mathbb{R}^{100}$ represents FLAME [23] expression parameters. $M_{1:T}$ denotes a T frame gesture clip, and $\hat{M}_{1:T}$ represents the corresponding reconstructed gesture clip. T is the number of the fixed frames during training. $M^b \in \mathbb{R}^{T\times 63}$ and $M^h \in \mathbb{R}^{T\times 90}$ represent body and hand poses. $E_{1:t} = \{e_1, ..., e_t\} \in \mathbb{R}^{t\times 64}$ denotes codebook vectors, and $Z_{1:t} = \{z_1, ..., z_t\} \in \mathbb{R}^{t\times 64}$ denotes latent vectors, where $t = \frac{T}{4} = 22$. $A_{1:T}$ is the MFCC (Mel-Frequency Cepstral Coefficients) feature of audio. id represents the speaker's identity drawn from a predefined set.

2.2. Face Generator

For the face, following [7], we adopt an encoder-decoder architecture, as illustrated in Figure 3. We learn face parameters through loss function as follows:

$$L_{face} = L_{jaw}(\theta_j, \hat{\theta}_j) + L_{expression}(\theta_e, \hat{\theta}_e), \qquad (1)$$

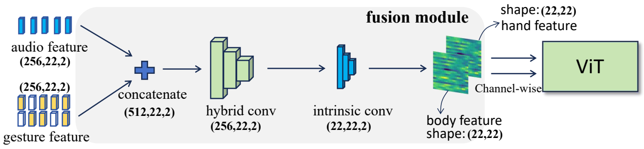

Fig. 4. Fusion module. We employ hybrid convolution to fuse audio and gesture features, enabling their interaction and forming a unified representation. Subsequently, intrinsic convolution downsamples the mixed features into the latent space.

where L_{jaw} and $L_{expression}$ are L_1 and L_2 reconstruction losses, resp. $\hat{\theta}$ denotes the corresponding reconstructed parameter.

2.3. Gesture Codebook

In the process of reconstructing and generating the hand and upper body parts, we use a **two-stage** approach. In the **first stage** we use VQ-VAE [24], which can learn a discrete representation and ensure that the poses are accurately reconstructed through latent vectors.

Our goal in this stage is to train VQ-VAE well enough to reconstruct pose data, preparing it for the generation task in the next stage. As in [7], we use a loss function L_{VQ}:

$$L_{VQ} = L_{rec}(M_{1:T}, \hat{M}_{1:T}) + L_{eq}(Z_{1:t}, E_{1:t}) + L_v(M_{1:T}, \hat{M}_{1:T}), \quad (2)$$

where L_{rec}, L_{eq} and L_v are reconstruction loss, codebook loss and velocity loss, resp.

2.4. Gesture Generator

The **second stage** focuses on generating high-fidelity gestures and enhancing model generalization. To save training time, we train the generator in the indice space of the codebook, with the loss function employing only **cross-entropy loss** as follows:

$$L_{cls} = CrossEntropy(I_{1:t}, ViT(fusion(\overset{m}{I}_{1:t}, a_{1:t}), id)), \quad (3)$$

where $I \in \mathbb{R}^{t \times 2}$ are the indices of poses. $fusion$ module is shown in Figure 4. $a_{1:t}$ denotes the audio feature after downsampling. $\overset{m}{I}_{1:t}$ denotes the indices of poses being masked.

To accelerate loss convergence, we incorporate the ground truth pose (GT) as an additional auxiliary input. Thus, our gesture generator is fed with $A_{1:T}$ and GT pose data $M_{1:T}$ in the training phase.

As shown in Figure 1, we first use two codebook encoders to get pose indices $I_{1:t}$ and use an audio encoder and 1×1 convolution to get audio feature $a_{1:t}$. Then, we apply a masking strategy (like [25]) to $I_{1:t}$ and pass the masked $I_{1:t}$ through an embedding layer to map $I_{1:t}$ and $a_{1:t}$ to the same dimensions. For pose features, we employ the DER strategy to enhance the model's learning capability by introducing random perturbations. DER applies dropout after the embedding layer, active during training but inactive during inference. This introduces noise to enhance robustness, and mitigates overfitting to improve generalization. Later, we fuse the audio and pose features using a fusion module and then input them into the ViT model. Notably, the input operation is carried out channel-wise, splitting the body and hand poses into two channels (injecting spatial information into the third dimension of the features). This is similar to an image with a shape of width × height × 2. We represent hand and body features as two channels within a single feature map (as shown in Figure 4), where each channel corresponds to one component yet is closely interconnected. Meanwhile, the width and height of the feature maps correspond to the temporal and spatial dimensions of the gesture clips respectively. For the temporal dimension, we do not perform compression, while for the spatial dimension, we conduct downsampling using intrinsic convolution. The integration and decoupling of spatio-temporal information at the feature map level enable the module to gain enhanced spatio-temporal understanding capabilities. Thus, channel-wise processing is crucial for the effectiveness of the RET module, as it is essential to fuse spatio-temporal information.

Method	*Diversity*↑	*FGD*↓	*MAE*↓	*BC*→
A and UN	8.4988	30.102	35.5114	0.8570
A and EN	9.3009	100.10	35.8544	0.8569
B and UN	8.2614	10.846	**35.4285**	0.8574
B and EN	8.9830	**2.4816**	35.9665	**0.8579**
C and UN	8.3830	16.744	35.4646	0.8567
C and EN	**10.971**	143.08	36.6753	0.8578

Table 1. **A** represents training with the Empty condition at a 10% probability. **B** denotes training with the dropout operation at a 10% probability. **C** represents using neither of the two strategies, i.e., not conducting any processing on the speech condition. **EN** represents enabling Equation 4 during inference, while **UN** represents not enabling.

In the ViT model, we first apply patchEmbed to different channels of the input data. Then, we add id and position encoding, both of which are fed into N ViT blocks (N is 15) to obtain the predicted indices $\hat{I}$. Finally, by applying the loss function to I and $\hat{I}$, we can train the model to converge.

2.5. Training Detail

After extensive experimentation, we find DER and masking strategies help to alleviate overfitting. These strategies introduce large random perturbations to the input data, not only guiding the model to learn more robust features, but also improving the generalization ability of the model. Additionally, we apply an Exponential Moving Average (EMA) technique during model training, which stabilizes the learning process in our supervised framework by maintaining a moving average of model parameters.

We use CFG [26] to train our model, but replace the Empty condition in Equation 3 with a dropout operation, which effectively enhances the performance of the model. During inference, we use the following Equation 4 in the last neural network layer before softmax to guide the generation process:

$$logit = s \cdot RET(a_{1:t}, id) - (s-1) \cdot RET(\phi, id), \quad (4)$$

where s is guidance scale, and $logit$ is the guided generation result. We can control the speaker's gesture style through id. In addition, we conduct experiments, as shown in Table 1, to prove that in our model, it is better to use dropout than the Empty condition during training.

3. INFERENCE PHASE

In the inference phase, we divide the process into face and gesture inference. For face inference, we use the VAE architecture from

	Diversity↑	*FGD*↓	*MAE*↓	*BC*→
ReCoM	**8.9830**	**2.4816**	35.966	**0.8579**
w/o CFG	8.2614	10.8462	35.428	0.8574
w/o IRI	8.7314	39.9367	**31.785**	0.8570
w/o EMA	8.1029	27.6172	35.436	0.8570
w/o DER	6.9025	146.394	35.295	0.8545
w/o masking	8.4321	71.0111	35.685	0.8560

Table 2. Ablation results on the SHOW dataset.

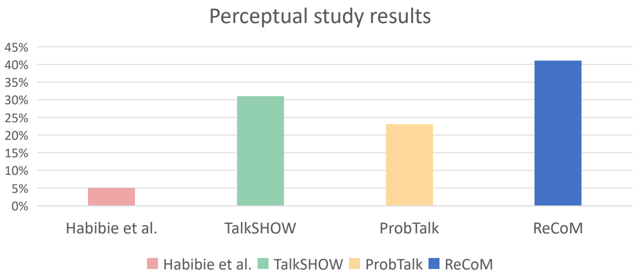

Fig. 5. We calculate the win rate of the evaluation.

section 2.2. Given audio input, the facial model trained on facial distributions generates a FLAME [23] result.

3.1. Iterative Reconstruction Inference

For the gesture inference part, we propose a novel inference strategy, which is termed IRI, to enhance the generation results, as shown in Figure 1. Specifically, the method initially takes speech features and fully masked motion indices as inputs. Then, the speech features and indices are repeatedly input into RET module to predict masked indices until all are recovered. In each iteration, the model automatically selects results that exceed the confidence threshold, while indices with results below the threshold are retained for re-prediction in the next iterations. The threshold is adaptively decreased in a linear manner in order to reduce the difficulty of data reconstruction. Making predictions in a fully non-chronological order helps alleviate the cumulative errors in the temporal sequence.

3.2. Temporal Smoothing Process

To generate long gesture sequences, we need to concatenate different result segments. We propose the Smoothing strategy as shown in Figure 1. Specifically, we divide the speech audio into several short segments in chronological order. For audio with a frame rate of 30 FPS, we split it into segments with 88 frames each. For segments S_i (excluding the first), we incorporate the last 8 frames of segment S_{i-1} into the segment S_i. This enables information from several independent segments to be transmitted in a temporal sequence.

4. EXPERIMENT

We train and test on the SHOW dataset with 27 hours data [7]. For generalization experiments, we test on BEAT2-English dataset with 26 hours data [8], without any fine-tuning.

Quantitative Comparisons. We compare ReCoM with [19], TalkSHOW [7] and ProbTalk [15]. We select a series of metrics to ensure a comprehensive and accurate assessment of the model's performance across multiple aspects. These metrics include **Diversity** [8], **FGD (Fréchet Gesture Distance)** [27, 28], **MAE (Mean Absolute Error)**, and **BC (Beat Consistency Score)** [29]. These metrics can effectively represent the performance of the gesture model. As shown in Table 3 and Table 4, the experimental results demonstrate that our model has high fidelity and good generalization ability.

Method	*Diversity*↑	*FGD*↓	*MAE*↓	*BC*→
GT	9.4850	0	0	0.8676
Habibie	7.5246	239.178	98.6942	0.9477
TalkSHOW	6.8678	66.1574	36.7540	**0.8713**
ProbTalk	7.6758	18.7028	36.0005	0.7837
ReCoM	**8.9830**	**2.4816**	**35.9665**	0.8579

Table 3. In-domain evaluation on **SHOW**. Downward (↓), upward (↑), and rightward (→) arrows indicate that lower, higher, and GT-closer values are better, respectively. Bold and underlined denote the best and second-best results.

Method	*Diversity*↑	*FGD*↓	*MAE*↓	*BC*→
GT	14.8500	0	0	0.8351
Habibie	7.5242	239.184	92.2333	0.9477
TalkSHOW	8.6990	98.3199	72.2534	0.8729
ProbTalk	8.2616	100.067	71.6509	0.8178
ReCoM	**11.1303**	**96.7793**	**71.5830**	**0.8469**

Table 4. Out-of-domain evaluation on **BEAT2**.

Perceptual Study. Objective metrics do not always reflect model performance. Therefore, to further verify the visual performance of ReCoM, we conduct a perceptual study. We generate a total of 81 visual videos for testing: 50 from the SHOW test set, and the remaining 31 are generated from wild TED audio. Twenty participants were asked to evaluate the videos generated by different methods. Each assessed all 81 test samples (presented in random order) and were asked to select their most preferred option. The results are presented in Figure 5.

Ablation Study. As shown in Table 2, to validate the effectiveness of our design, we conduct ablation experiments by removing our improvement strategies. Although our ReCoM is not the best in all metrics, we choose it because FGD has the biggest impact on visual effect. Spectral analysis using short-time Fourier transform indicates that IRI's increase in MAE is concentrated in high-frequency bands and that IRI preserves, or even slightly improves, accuracy at low and mid frequencies. This shows that IRI functions by selectively smoothing high-frequency micro-motions. Experiments with exponential and cosine decay thresholding yield results comparable to linear decay, implying that IRI's effectiveness depends more on the iterative process itself than on the thresholding strategy.

5. CONCLUSION

We design RET by retaining the structural characteristics of ViT, and enable the model to process spatio-temporal information through channel-wise operations. Additionally, we apply some strategies to improve model learning, and use a new inference strategy to enhance the generation results. Our ReCoM achieves impressive results, not only improving fidelity but also enhancing generalization.

6. REFERENCES

[1] Han Liang, Jiacheng Bao, Ruichi Zhang, Sihan Ren, Yuecheng Xu, Sibei Yang, Xin Chen, Jingyi Yu, and Lan Xu, "Omg: Towards open-vocabulary motion generation via mixture of controllers," in *CVPR*, 2024.

[2] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng, "Momask: Generative masked modeling of 3d human motions," in *CVPR*, 2024.

[3] Xu Shi, Chuanchen Luo, Junran Peng, Hongwen Zhang, and Yunlian Sun, "Fg-mdm: Towards zero-shot human motion generation via chatgpt-refined descriptions," in *ICPR*, 2024.

[4] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter, "Listen, denoise, action! audio-driven motion synthesis with diffusion models," *ACM Trans. Graph.*, July 2023.

[5] Wei Yao, Yunlian Sun, Hongwen Zhang, Yebin Liu, and Jinhui Tang, "Hosig: Full-body human-object-scene interaction generation with hierarchical scene perception," in *AAAI*, 2026.

[6] Jinming Zhang, Yunlian Sun, Hongwen Zhang, and Jinhui Tang, "Edmg: Towards efficient long dance motion generation with fundamental movements from dance genres," in *ACM MM*, 2025.

[7] 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 *CVPR*, 2023.

[8] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black, "Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling," in *CVPR*, 2024.

[9] 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*. Springer, 2022.

[10] Xingqun Qi, Hengyuan Zhang, Yatian Wang, Jiahao Pan, Chen Liu, et al., "Cocogesture: Toward coherent co-speech 3d gesture generation in the wild," *Information Fusion*, 2025.

[11] William Peebles and Saining Xie, "Scalable diffusion models with transformers," in *ICCV*, October 2023, pp. 4195–4205.

[12] Tenglong Ao, Zeyi Zhang, and Libin Liu, "Gesturediffuclip: Gesture diffusion model with clip latents," *ACM Transactions on Graphics (TOG)*, 2023.

[13] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu, "Taming diffusion models for audio-driven co-speech gesture generation," in *CVPR*, June 2023.

[14] Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen, "Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation," in *CVPR*, 2024.

[15] Yifei Liu, Qiong Cao, Yandong Wen, Huaiguang Jiang, and Changxing Ding, "Towards variable and coordinated holistic co-speech motion generation," in *CVPR*, 2024.

[16] Wei Yao, Hongwen Zhang, Yunlian Sun, and Jinhui Tang, "Staf: 3d human mesh recovery from video with spatio-temporal alignment fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[17] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu, "Pymaf-x: Towards well-aligned full-body model regression from monocular images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[18] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun, "Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop," in *ICCV*, October 2021, pp. 11446–11456.

[19] 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*, 2021.

[20] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[21] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, 2014.

[22] Yarin Gal and Zoubin Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," in *Neural Information Processing Systems*, 2015.

[23] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero, "Learning a model of facial shape and expression from 4d scans.," *ACM Trans. Graph.*, 2017.

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

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019*, Jill Burstein, Christy Doran, and Thamar Solorio, Eds., Minneapolis, Minnesota, June 2019.

[26] Jonathan Ho and Tim Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[27] 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)*, 2020.

[28] Taras Kucherenko, Pieter Wolfert, Youngwoo Yoon, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter, "Evaluating gesture generation in a large-scale open challenge: The genea challenge 2022," *ACM Transactions on Graphics*, 2023.

[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 *CVPR*, 2022.