SpeechCAT: Cross-Attentive Transformer for Audio to Motion Generation

Sebastian Deaconu
Delft University of Technology
Delft, Netherlands
S.Deaconu@student.tudelft.nl

Xiangwei Shi
Delft University of Technology
Delft, Netherlands
X.Shi-3@tudelft.nl

Thomas Markhorst
Delft University of Technology
Delft, Netherlands
T.C.Markhorst@tudelft.nl

Jouh Yeong Chew
Honda Research Institute Japan
Saitama, Japan
jouhyeong.chew@jp.honda-ri.com

Xucong Zhang
Delft University of Technology
Delft, Netherlands
xucong.zhang@tudelft.nl

2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI) | 979-8-3503-7893-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/HRI61500.2025.10974020

***Abstract*—Audio-to-motion generation is an important task with applications in virtual avatar creation for XR systems and intelligent robot control in daily life scenarios. However, most existing motion generation methods rely on a single encoder-decoder architecture to model all body parts simultaneously, which limits their ability to capture the diverse and complex motions exhibited by humans. In this paper, we propose a novel method, SpeechCAT, that employs three separate encoder-decoder modules to individually model the motions of the face, body, and hands. To capture the relationships and synchronization among these body parts, we introduce a cross-attention mechanism to effectively learn their correlations. SpeechCAT ensures sufficient capacity to model the unique characteristics of each body part while preserving the coherence between them. Our experimental results demonstrate the superiority of SpeechCAT over baseline methods, highlighting its effectiveness in generating diverse, realistic, and synchronized motions with face, body, and hand parts.**

***Index Terms*—Audio-to-motion Generation, Multi-modal**

I. Introduction

Given a human speech audio input, motion generation aims to synthesize synchronized human motion corresponding to the audio like another human being movements. This foundational technique has broad applications, including film production [1], virtual digital avatar generation [2], and human-robot interaction [3]. Consequently, this area of research has garnered significant attention, with numerous studies contributing to advancements in the field [4]–[7]. Specifically, the generated motion can be directly used to drive the intelligent robot to interact naturally with the human user.

Early works in audio-to-motion generation mainly focused on single motion domains, such as facial motion [8], hand gestures [9], or 3D body skeletons [10]–[12]. Although these studies achieved promising progress, they often lack the integration of multiple body parts, limiting their applicability to comprehensive human motion generation. Recent studies [13]–[15] have begun addressing this limitation by leveraging holistic information to generate full-body human motion from speech. These methods integrated multiple body parts, including face, body, and hands, leveraging human mesh representation [16], [17] as the intermediate or final outputs.

Human motion is a continuous and dynamic process, making it computationally expensive to model. To address this challenge, the Vector Quantized Variational Autoencoders (VQ-VAE) framework [18] has been applied in generating human facial [19]–[21], body [22], and hand motions [13]. This approach effectively captures discrete representations of continuous motion, significantly reducing the computational burden while preserving motion quality. As for facial motions, these methods either adopted a lip regressor [15] to synthesize synchronized lip movements from audio or trained an encoder-decoder model to produce synchronized mouth motions in a deterministic manner [13].

While these motion generation methods have demonstrated promising results, they overlooked the correlations among the three body parts. For instance, hand gestures are closely correlated with wrist positions, which, in turn, should affect forearm movement. Even though face motions, especially lip movements, are linked to the audio input directly, isolating the face domain from the body and hands hinders the generation of natural, synchronized human motions. This limitation shows the need to model across body parts correlations explicitly to enhance the realism and coherence of generated motion.

To address this limitation, we propose a novel method, SpeechCAT, for audio-to-motion generation. Different from previous works, SpeechCAT includes three VQ-VAE to encode three discrete codebooks for the face, body, and hands, respectively. Once the codebooks are fully trained, we concatenate the encoded audio features with body-hand-face triplet indices from the codebooks, which are then processed by a cross-attention transformer module. To model the correlation and synchronization across different body parts of face, body, and hands, we perform the cross-attention across the three body parts. By this way, SpeechCAT explicitly captures and models the correlations and synchronization among the face, body, and hands, ensuring the generated motions are coherent and realistic across all domains.

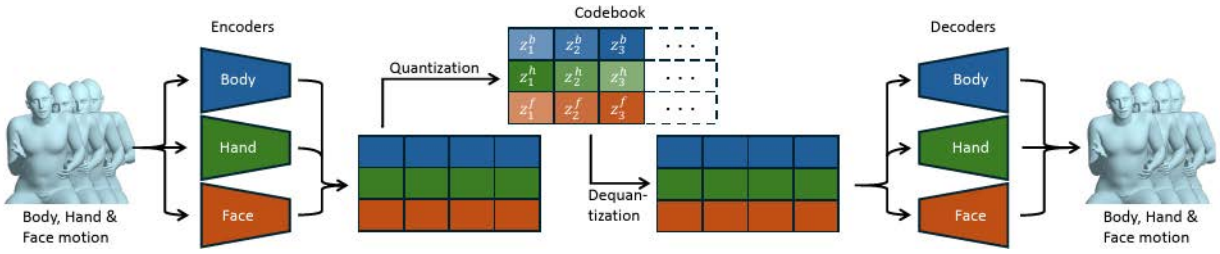

Fig. 1. VQ-VAE module that learns the discrete codebook indexes by reconstruction of the input motion. Different from previous works, we propose to separate the body motions into three codebooks for face $\mathcal{Z}^f$, body $\mathcal{Z}^b$, and hand $\mathcal{Z}^h$, respectively. Each body part has its own encoder and decoder.

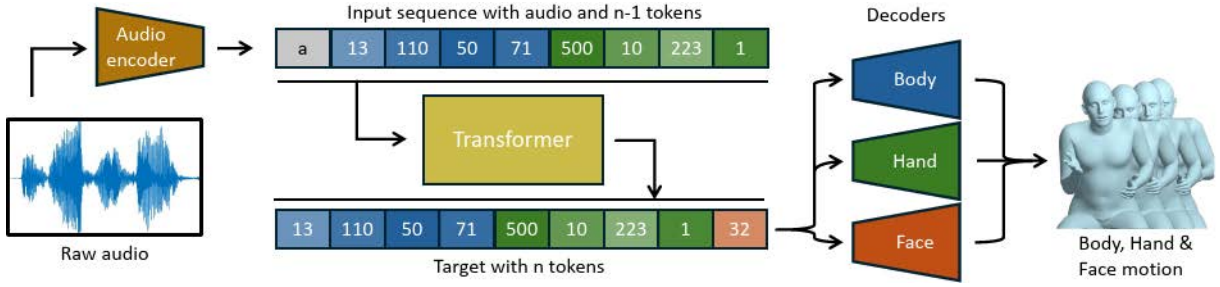

Fig. 2. Overview of the motion generation module that takes the audio as input to generate the corresponding body motion. It first maps the audio feature to the motion indexes learned in the VQ-VAE. The core of the motion generator is a transfer to predict the motion index in an auto-regressive manner. The motion index is mapped into motion by the decoder of the VQ-VAE.

Our experimental results demonstrate that SpeechCAT significantly outperforms baselines that use a single VQ-VAE or without the cross-attention module, highlighting its effectiveness in producing synchronized, diverse, and natural human motion. Especially, we find that the SpeechCAT can provide better synchronization in the generated motions.

II. Method

The proposed SpeechCAT takes the audio as input to generate the face, body, and hand movements. The proposed framework comprises two modules including a VQ-VAE encoder-decoder and a cross-attentive transformer. The VQ-VAE creates a mapping from motion sequences to discrete cookbooks for face, body, and hand parts. The transformer acts as a motion generator that takes audio as input to predict the index of the codebook, which can then be decoded to motion with the decoder from the VQ-VAE.

A. Motion VQ-VAE

The overview of the VQ-VAE is illustrated in Figure 1. It reconstructs the input motion to perform unsupervised feature learning. These features are constructed as codebooks. Specficially, to train the VQ-VAE for motions, we first define sequence motions of face $M^f = (m_1^f, ..., m_T^f) \in \mathbb{R}^{103\times T}$, body $M^b = (m_1^b, ..., m_T^b) \in \mathbb{R}^{63\times T}$, and hand $M^h = (m_1^h, ..., m_T^h) \in \mathbb{R}^{63\times T}$, where T is the number of frames. The dimensions of each part are based on the SMPL-X model [16]. We encode these motions to the codebook $\mathcal{Z} = \{z_i\}_{i=1}^{|\mathcal{Z}|}$. which contains discrete motion vectors $z_i \in \mathbb{R}^{d_z}$ that represent a quantized latent space.

Different from previous method [13], which uses a shared codebook, we define separate codebooks for the face $\mathcal{Z}^f = \left\{z_i^f\right\}_{i=1}^{|\mathcal{Z}^f|}$, body $\mathcal{Z}^b = \left\{z_i^b\right\}_{i=1}^{|\mathcal{Z}^b|}$, and hand $\mathcal{Z}^h = \left\{z_i^h\right\}_{i=1}^{|\mathcal{Z}^f|}$, respectively, where the motion vectors are defined as $z_i^b, z_i^h, z_i^f \in \mathbb{R}^{64}$. This separation ensures that each body part—face, body, and hands—has sufficient latent space to capture its unique motion characteristics. To model the correlations between these parts, we employ a cross-attention mechanism that integrates the information from their respective latent spaces. Specifically, we use three separated encoders to map the motion of the face M^f, body M^b, and hand M^h into three different latent spaces for face $\mathcal{Z}^f$, body $\mathcal{Z}^b$, and hand $\mathcal{Z}^h$ by quantization on the feature space. We then decode the latent space to the feature space and then decode these features back to the human motions.

Optimization goal. The VQ-VAE is optimized by the loss component $\mathcal{L}_{VQ}$ [18] which is composed of a reconstruction loss $\mathcal{L}_{rec}$, an embedding loss $\mathcal{L}_{embed}$ and a commitment loss $\mathcal{L}_{commit}$, which can be written as the following:

$$\mathcal{L}_{VQ} = \mathcal{L}_{rec}(M_{1:T}, \hat{M}_{1:T}) + \underbrace{\|\text{sg}[E_{1:T}] - Z_{1:T}\|}_{\mathcal{L}_{embed}} + \beta \underbrace{\|E_{1:T} - \text{sg}[Z_{1:T}]\|}_{\mathcal{L}_{commit}} \quad (1)$$

The $\mathcal{L}_{rec}$ is the MSE reconstruction loss, sg is the stop gradient operator for the codebook embedding loss and β is the trade-off hyperparameter for the commitment loss. We

combine the $\mathcal{L}_{VQ}$ for face M^f, body M^b, and hand M^h together as the loss for the motion VQ-VAE training.

B. Motion Generator

The overview of the motion generation is shown in Figure 2. The input audio x is first encoded with an audio encoder from wav2vec [23] and then is mapped to the late feature vector $\mathcal{Z}$ that learned from the VQ-VAE. Note the late space $\mathcal{Z}$ contains face $\mathcal{Z}^f$, body $\mathcal{Z}^b$, and hand $\mathcal{Z}^h$ components. Essentially, the audio encoder is a classifier that predicts which class from the quantize a latent value from the codebook $\mathcal{Z}$. We formulate the motion latent space prediction process as an autoregressive task [24]. Given a speech input and the previously predicted motion vector in $\mathcal{Z}$, we use a transformer to predict the distribution of the next possible indexes of the motion vector in $\mathcal{Z}$. The predicted latent space features can be decoded back to a motion through that trained decoder from the VQ-VAE. At inference, we start from the speech token and generate the next motion tokens in an auto-regressive manner, until the sequence is equal in size to the initial speech sequence.

Optimization goal. To optimize the motion generator, we maximize the log-likelihood of our data distribution. Given that our sequence data likelihood can be mapped as $p(\mathcal{Z} \mid x) = \prod_{i=1}^{|\mathcal{Z}|} p(\mathcal{Z}_i \mid x, \mathcal{Z}_{i-1})$, we then have the following Transformer loss:

$$\mathcal{L}_{\text{trans}} = \mathbb{E}_{\mathcal{Z}\sim p(\mathcal{Z})}\left[-\log p(\mathcal{Z} \mid x)\right] \tag{2}$$

Cross-Attention. In addition to causal self-attention, To enable the comprehensive communication and synchronization between different body parts i.e., body, hands, and face, SpeechCAT employs a cross-attention module formulated as follows:

$$\text{CrossAttention} = \text{Softmax}\left(\frac{Q_{\text{part1}} K_{\text{part2}}^T}{\sqrt{d_k}}\right) \tag{3}$$

where $Q_{\text{part1}} \in \mathbb{R}^{t\times d}$ and $K_{\text{part2}} \in \mathbb{R}^{t\times d}$ represent the query and key matrices of paired two different body parts i.e., hands and body, face and hand, or hand and face. This cross-attention mechanism allows each body part to attend to other parts, effectively sharing information to improve cohesiveness in the generated motion. Each body part has its unique query, key, and value representations. Thus, the attention is directed across body parts rather than within a single part, ensuring motion that reflects coordinated gestures or body-language cues.

There is a discrepancy in index quality between inference and training. While in training all $i-1$ indexes are assumed to be correct, there is no guarantee that the indexes used for generation are relevant for the conditional probability. To address this issue, we replace $\alpha\times 100\%$ of ground truth indexes as explained in [22]. This combined with index sampling from the predicted distributions ensures diversity for our motion generator. For the trade-off of stability, α was set to 0.4.

III. Experiment

To evaluate the ability of SpeechCAT, we conduct experiments on the TalkSHOW dataset [13]. Specifically, an 80%/10%/10% train-val-test split on the dataset is used and videos are between three and four seconds. We compare our SpeechCAT with two baselines. The first baseline ***One-code*** represents the method that uses only one latent space for face, body, and hand motions. Another baseline ***w/o cross-att.*** represents our method without the cross-attention module between different body parts.

Method	Body & Hand			Face	
	L2 ↓	Diversity ↑	MSD ↓	L2 ↓	LVD ↓
w/o cross-att.	12.47	0.37	0.1532	0.2039	0.0314
One-code	12.007	0.32	0.1530	**0.1892**	0.0321
SpeechCAT	**11.7**	**0.43**	**0.1519**	0.2048	**0.0312**

TABLE I
COMPARISON OF PROPOSED SPEECHCAT AND TWO BASELINES WITH MULTIPLE ERROR METRICS. WE SEPARATE THE BODY & HAND AND FACE SINCE THESE JOINTS ARE IN DIFFERENT MAGNITUDES.

A. Experimental setup

Evaluation metrics: Even though we use a generative, non-deterministic Transformer model, we test its ability to learn an informative and robust feature space, thus also using metrics such as coherence, synchronization, and stability. The full list of metrics used is the following:

- **L2 error**: L2 distance between ground-truth and generated joints positions and reported as the average error in millimeters.
- **LVD**: Landmark Velocity Difference calculates the velocity difference between ground-truth and generated joints. It specifically measures the synchronization between the speech and the generated motion, thus, we only use this metric on the face part.
- **Diversity**: We calculate variance across 16 samples of body and hand motions for the same audio input. Since diversity mainly exists in the body and hand, we do not use the metric for the face region.
- **MSD**: Mean Squared Displacement measures the average frame-to-frame displacement of joints, thus quantifying temporal smoothness.
- **BPSS**: Body Part Synchronization Score is a weighted average of the following motion metrics, calculated across body parts including body, face, and hands:
  - **Cross-Correlation**: Computes the correlation between different body parts to measure general alignment.
  - **Mean Phase Coherence**: Measures phase alignment to capture synchronization in motion cycles.
  - **Velocity and Acceleration Alignment**: Cosine similarity to capture alignment in the direction and speed of movement between body parts.

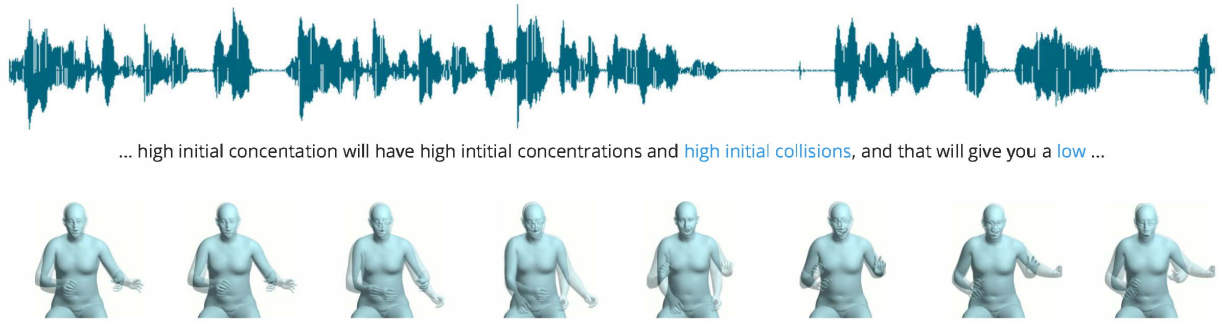

Fig. 3. A demonstration of the synchronization between the input audio and motion generated by SpeechCAT. The method generates motion consistent with the rhythm and tone of input audio. The intonation of the strengthening words "high" and "low" is directly expressed through the hand motions.

Model Ablation Study				
Model	**corr ↑**	**mpc ↑**	**vel_al ↑**	**acc_al ↑**
W/o cross-attention	0.429	0.623	0.853	2.674
One-code	0.349	0.577	0.844	2.554
SpeechCAT	**0.618**	**0.732**	**0.881**	**2.761**

TABLE II

PERFORMANCES ON THE MOTION GENERATION SYNCHRONIZATION ACROSS THREE BODY PARTS. METRICS INCLUDE CORRELATION (CORR), MEAN PHASE COHERENCE (MPC), VELOCITY ALIGNMENT (VEL_AL), AND ACCELERATION ALIGNMENT (ACC_AL).

B. Motion Generation Analysis

We compare our SpeechCAT with two baselines in Table I, evaluating the body & hand and face parts. From the table, it is evident that SpeechCAT outperforms the two baselines across most error metrics, with the exception of the L2 error for the face part. This result highlights the advantages of the proposed SpeechCAT with the cross-attention module and multiple encoder-decoder modules architecture.

Notably, SpeechCAT demonstrates significant improvements in diversity in body and hand parts, a critical metric for the audio-to-motion generation task. The diversity performance drops substantially for the One-code baseline, underscoring the limitation of using a single encoder-decoder architecture. For the L2 error on the body and hand parts, the baseline without the cross-attention module exhibits the worst performance compared to the other methods. This emphasizes the necessity of the cross-attention mechanism in SpeechCAT for effectively modeling the correlations between body and hand motions. SpeechCAT does not achieve substantial improvements for the face region. This could be attributed to the relatively isolated nature of facial motions, which are less influenced by body and hand movements.

Overall, we conclude that the proposed SpeechCAT effectively maps the three body parts such as face, body, and hands into separate latent spaces, while leveraging cross-attention to model their correlations. This design ensures improved motion diversity, stability, and robustness, further enhanced by the transformer-based motion generator.

C. Synchronization Analysis

For the audio-to-motion generation task, aligning different body parts is crucial. To evaluate the synchronization among the face, body, and hand motions, we show the results in Table II using the BPSS metrics. The results show that the proposed SpeechCAT significantly outperforms the two baselines across all error metrics. Notably, SpeechCAT achieves substantial improvements in cross-correlation and mean phase coherence, which are specifically designed to assess the alignment and correlation among different body parts. These improvements demonstrate the effectiveness of our SpeechCAT in modeling the relationships between the face, body, and hands. In contrast, the One-code baseline yields the poorest performance across all metrics, highlighting the limitations of using a single shared latent space for all body parts. This underscores the advantage of employing separate latent spaces tailored to each body part, as proposed in SpeechCAT. Overall, the combined approach of separate latent spaces integrated with cross-attention in SpeechCAT not only enhances robustness but also ensures better coherence and synchronization across the body parts compared to using standalone VQ-VAEs.

D. Qualitative Analysis

An example of generated full-body motion can be seen in Figure 3. For the expression "high initial collisions" which is followed by a pause in speech to attract the listener's attention, the model generates a body that lifts its hands in front with open palms. A similar motion is generated for the word "low" which is also used as a strengthening tone. The hands are lifted beforehand and slowly put down to accentuate the word. This is a natural expression that complements the intonation as well as the rhythm of the audio.

IV. CONCLUSION

Our approach models the face, body, and hands separately using a three-encoder-decoder architecture, capturing the unique variations in motion for each body part by mapping them into distinct latent spaces. To account for the correlation between these body parts, we introduce a cross-attention mechanism that learns and integrates the correlations among different body parts. Experimental results demonstrate the effectiveness of the proposed architecture in generating diverse and coherent motions across all body parts.

REFERENCES

[1] J. Lee, B. Han, and S. Choi, "Motion effects synthesis for 4d films," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 10, pp. 2300–2314, 2015.

[2] B. Zhang, Y. Cheng, C. Wang, T. Zhang, J. Yang, Y. Tang, F. Zhao, D. Chen, and B. Guo, "Rodinhd: High-fidelity 3d avatar generation with diffusion models," in *Eur. Conf. Comput. Vis.*, vol. 2, 2024.

[3] J. Bütepage, H. Kjellström, and D. Kragic, "Anticipating many futures: Online human motion prediction and generation for human-robot interaction," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 4563–4570.

[4] N. Kolotouros, T. Alldieck, E. Corona, E. G. Bazavan, and C. Sminchisescu, "Instant 3d human avatar generation using image diffusion models," 2024.

[5] W. Zhu, X. Ma, D. Ro, H. Ci, J. Zhang, J. Shi, F. Gao, Q. Tian, and Y. Wang, "Human motion generation: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[6] G. E. Henter, S. Alexanderson, and J. Beskow, "Moglow: Probabilistic and controllable motion synthesis using normalising flows," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–14, 2020.

[7] V. Korzun, A. Beloborodova, and A. Ilin, "The finemotion entry to the genea challenge 2023: Deepphase for conversational gestures generation," in *Proceedings of the 25th International Conference on Multimodal Interaction*, 2023, pp. 786–791.

[8] E. Ng, S. Subramanian, D. Klein, A. Kanazawa, T. Darrell, and S. Ginosar, "Can language models learn to listen?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 083–10 093.

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

[10] M. H. Mughal, R. Dabral, I. Habibie, L. Donatelli, M. Habermann, and C. Theobalt, "Convofusion: Multi-modal conversational diffusion for co-speech gesture synthesis," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 1388–1398.

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

[12] S. Qian, Z. Tu, Y. Zhi, W. Liu, and S. Gao, "Speech drives templates: Co-speech gesture synthesis with learned templates," in *Int. Conf. Comput. Vis.*, 2021, pp. 11 077–11 086.

[13] H. Yi, H. Liang, Y. Liu, Q. Cao, Y. Wen, T. Bolkart, D. Tao, and M. J. Black, "Generating holistic 3d human motion from speech," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 469–480.

[14] M. Sun, C. Xu, X. Jiang, Y. Liu, B. Sun, and R. Huang, "Beyond talking–generating holistic 3d human dyadic motion for communication," *arXiv preprint arXiv:2403.19467*, 2024.

[15] E. Ng, J. Romero, T. Bagautdinov, S. Bai, T. Darrell, A. Kanazawa, and A. Richard, "From audio to photoreal embodiment: Synthesizing humans in conversations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1001–1010.

[16] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 10 975–10 985.

[17] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 851–866.

[18] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Adv. Neural Inform. Process. Syst.*, vol. 30, 2017.

[19] J. Wang, K. Zhao, S. Zhang, Y. Zhang, Y. Shen, D. Zhao, and J. Zhou, "Lipformer: High-fidelity and generalizable talking face generation with a pre-learned facial codebook," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 13 844–13 853.

[20] Y. Lu, J. Chai, and X. Cao, "Live speech portraits: real-time photorealistic talking-head animation," *ACM Transactions on Graphics (ToG)*, vol. 40, no. 6, pp. 1–17, 2021.

[21] E. Ng, H. Joo, L. Hu, H. Li, T. Darrell, A. Kanazawa, and S. Ginosar, "Learning to listen: Modeling non-deterministic dyadic facial motion," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 20 395–20 405.

[22] J. Zhang, Y. Zhang, X. Cun, Y. Zhang, H. Zhao, H. Lu, X. Shen, and Y. Shan, "Generating human motion from textual descriptions with discrete representations," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 14 730–14 740.

[23] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Adv. Neural Inform. Process. Syst.*, vol. 33, pp. 12 449–12 460, 2020.

[24] A. Stein, S. Sharpe, D. Bergman, S. Kumar, C. B. Bruss, J. Dickerson, T. Goldstein, and M. Goldblum, "A simple baseline for predicting events with auto-regressive tabular transformers," *arXiv preprint arXiv:2410.10648*, 2024.