Voicing Your Emotion: Integrating Emotion and Identity in Cross-Modal 3D Facial Animations

Category: Research

Paper Type: Algorithm/technique

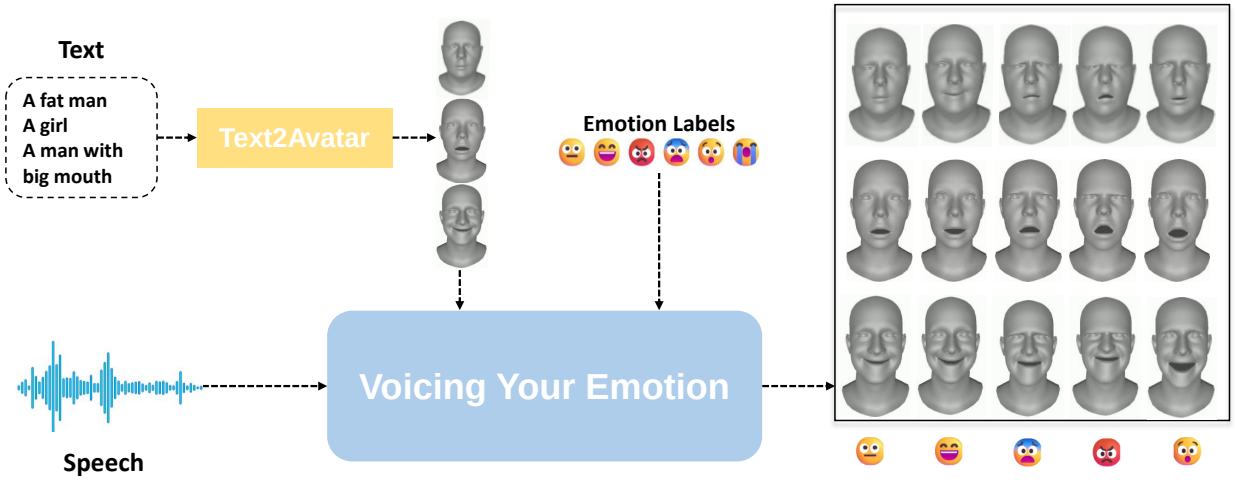


Fig. 1: Our method can combine text descriptions with audio features to generate accurate 3D facial animations rich in emotions. Firstly, we generate the corresponding 3D facial template from the text based on our designed Text2Avatar, and then combine it with audio and emotional labels to generate emotionally rich results.

Abstract— Recent advances in speech-driven 3D facial animation have yielded promising results, yet capturing intricate expressiveness, particularly in emotion and identity, remains an intricate challenge. While conventional studies often prioritize lip synchronization, they frequently overlook the full spectrum of emotional nuances and the uniqueness of individual identity. Addressing this deficiency, we unveil a pioneering method tailored to produce 3D facial expressions that resonate deeply with emotion and identity, guided by speech and user-provided prompt words. Our key insight is an emotion-identity fusion mechanism, a pre-trained self-reconstruction codebook, meticulously crafted from a wide array of emotional facial movements, serving as an expressive motion benchmark. Using this foundation, prompt words are seamlessly transformed into facial representations that capture both emotion and identity, and are then adeptly projected onto 3D templates. When synergized with speech audio and a user-specified emotion, our algorithm breathes life into a 3D avatar, reflecting the desired emotion and identity. The model's efficacy is bolstered by an advanced autoregression technique, marrying emotion and identity through an innovative feature fusion mechanism and a purpose-built loss function. Consequently, our approach emerges as a robust tool for crafting 3D talking avatars, rich in emotional depth and distinctive identity.

1 Introduction

The advancement of 3D facial animation sits at an intricate nexus where technology grapples with the profound complexities of human emotional and identitarian expressions. Scholarly undertakings, until now, have focused on achieving remarkable strides in realizing photorealistic animations and achieving precise lip-synchronization as underscored in notable works [11, 17, 21, 22, 27, 28, 30–32, 41, 46]. However, a discernible lacuna persists – the integration of emotional expressiveness and distinct identity, elements that transform technical renderings into lifelike, expressive avatars.

Our exploration is situated within this context, acknowledging the unprecedented challenges inherent in the fusion of emotion and identity. The dynamic nature of emotions, characterized by their transient and multifaceted expressions, poses a significant challenge. Emotions are not static; they are fluid, evolving, and context-dependent, rendering their capture and representation a complex endeavor.

Similarly, identity, imbued with both universal and individual characteristics, adds another layer of complexity. The uniqueness of individual identity, shaped by an intricate tapestry of cultural, social, and personal nuances, demands a highly sophisticated approach to ensure authenticity in representation within the animation.

In our pursuit to address these complexities, we unveil a methodological innovation epitomized in the emotion and identity feature fusion module. This initiative is supported by an analytical rigor applied to enriched datasets including VOCASET [11] and BIWI [18], and enhanced by the sophisticated EMOCA [12] method and the MEAD [42] dataset.

Our approach is not without its challenges. Emotions, with their inherent ephemerality and variability, require nuanced algorithms capable of capturing and translating these transient states into discernible, realistic expressions. Identity, on the other hand, brings forth the challenge of representing the uniqueness of individual traits without compromising the universal aspects that are shared across different individuals.

Despite these challenges, our fusion module stands as a testament to the intricate dance between rigorous technical innovation and the elusive, yet profound, aspects of human expressivity. We articulate a realm where 3D facial animations are not mere visual and auditory renderings but are imbued with emotional and identitarian nuances, making each piece a rich, multifaceted entity echoing the complexities of human communication.

Our method shows the advantages of achieving rich emotion and identity expression. The contributions of our work are as follows:

- Our endeavor to bridge the gap between textual prompts and 3D

physical states is encapsulated in the proposal of a cross-modality 3D talking face generation method. We are transitioning from static, one-dimensional animations to dynamic, multifaceted representations that breathe life into character faces with diverse styles and expressions. This innovation transcends the conventional boundaries, offering a customized, flexible, and adaptive solution that resonates with the complexity of real-world scenarios.

- Rooted in the premise of integrative expressiveness, we introduce an emotion and identity feature fusion module. This is not just a technological advancement but a creative leap, ensuring that each animation is a true reflection of specified speaker emotions. It has the adept capability to amalgamate diverse speech styles of characters, creating a rich tapestry of expressions that are as dynamic and varied as the human experience itself.
- A new loss function. An effective loss function method focuses on the overall face, which can increase the amplitude of changes in the speaker's mouth with accurate mouth shape and make the overall face more dynamic.

2 Related Works

2.1 Speech-Driven 3D Facial Animation

The progression in the field of 3D facial animation, especially in the context of audio-visual mapping, has been marked by notable contributions, each addressing specific facets of this complex domain. MeshTalk [32] is a commendable attempt at disentangling audio-correlated and uncorrelated facial information, employing a categorical latent space. Nevertheless, it is constrained by the suboptimal expressiveness of its chosen latent space, which manifests as unstable animation quality, particularly in settings where data is limited.

VOCA [10] marks another significant contribution, distinguished by its utilization of advanced audio feature extraction models. It excels in generating a diverse range of facial animations that correspond adeptly to a variety of speaking styles. In contrast, FaceFormer [17] leverages the Transformer architecture [40] to ensure temporal stability in animations by encompassing an extended audio context. However, both VOCA and FaceFormer are impeded by a shared challenge: they directly tackle the ill-posed nature of audio-visual mapping, characterized by significant uncertainty and ambiguity. This direct approach precipitates an over-smoothing problem, curtailing the vibrancy and dynamism of the generated animations.

Building on these foundations, CodeTalker [46] enhances the framework introduced by FaceFormer. It integrates a VQ-VAE [39] motion prior that has been separately trained. This innovation introduces discrete motion primitives for facial motion representation in a rich contextual manner, demonstrating heightened efficacy in learning general priors.

However, a prevailing limitation across these studies remains the predominant focus on the precision of lip movements. The resultant animations, while technically adept, often lack the controllable and vivid expressions crucial for lifelike representations. There's a discernible gap in the expressiveness of the animations, indicative of a broader need for methods that amalgamate technical accuracy with the nuanced incorporation of complex, dynamic human expressions.

Our research is situated within this context, addressing the expressiveness deficit and aiming to elevate the field from mere technical accuracy to an enriched domain where animations are characterized by their emotional and identitarian authenticity.

Due to the extensive and easily accessible 2D audio-visual dataset [1, 2, 9, 19, 35, 42, 44, 49], 2D speaking faces [3, 5, 7, 8, 14, 15, 20, 23, 25, 38, 43, 44] have produced satisfactory results. In the field of 3D talking faces [6, 11, 17, 21, 26, 30–32, 36, 46], deep learning based methods largely rely on a large amount of data, utilizing supervised methods [11, 17, 21, 32, 37, 46, 48] to generate highly synchronized speaking faces with lips. Previous work often used VOCASET, BIWI and Multiface [45] as training sets. These datasets are limited in richness of emotion and speaking styles. VOCASET face mesh register with FLAME [24] topology. Flame is a universal model for 3D faces. The input of a trained FLAME model is a parameter vector, including shape parameters, posture parameters, and expression parameters. These parameters control the identity features of the face, the rotation and translation of the head, and the expression changes of the face, respectively. The output of the model is a three-dimensional facial mesh, which has a set of vertices (nv=5023) and faces (nv=9976). This 3D facial grid can be used to represent different facial shapes and poses. Due to its excellent plasticity and ability to convey rich facial features, our work drives the FLAME model to generate speaking faces.

2.2 Adaptation to 3D from Rich 2D Audio-Visual Datasets

The abundance and accessibility of 2D audio-visual datasets [1, 2, 9, 35, 42, 49] have paved the way for significant advancements in the generation of 2D speaking faces [3, 5, 7, 8, 14, 15, 20, 23, 25, 38]. These datasets have been instrumental in achieving commendable results in this domain.

Transition to 3D Speaking Faces and the Associated Challenges However, when transitioning into the sphere of 3D talking faces [11, 17, 21, 30–32, 46], the reliance on voluminous datasets becomes apparent. The deep learning methodologies in current paradigms predominantly employ supervised methods [11, 17, 21, 32, 37, 46], aiming to generate speaking faces exhibiting high synchronization, particularly in the lip movement.

Noteworthy contributions have often utilized VOCASET, BIWI, and Multiface [45] as their foundational training sets. However, a critical evaluation reveals a marked limitation – these datasets exhibit a lack of diversity in emotional expressiveness and speaking styles.

VOCASET's face mesh, registered with FLAME topology, emerges as a pivotal tool. FLAME stands as a universal model adept at the intricate rendering of 3D faces. It operates through an input parameter vector, encompassing shape, posture, and expression parameters, which intricately dictate the identity, rotation and translation, and expressive changes of the face respectively. The model's output, a sophisticated three-dimensional facial mesh, is characterized by a set of vertices (nv=5023) and faces (nv=9976), capable of representing an array of facial shapes and poses.

In the context of these existing tools and recognized limitations, our work adopts and extends the FLAME model to generate speaking faces. We recognize FLAME's exemplary malleability and its capability to represent rich facial features. By leveraging this model, we aim to transcend the identified gaps, offering an innovative pathway to more expressive and dynamically responsive 3D talking faces.

2.3 Identity Customization and Emotional Control

There are few methods to provide users with options for identity customization and emotion control of generated 3D avatars. MeshTalk can generate multiple results for the same audio input, but there is no mechanism that allows for any control over emotions. FaceFormer and CodeTalker control speech style by inserting training individuals' style vectors, which cannot achieve simple emotional editing. EmoTalk [29] is a method of animating emotional 3D faces through voice input, which requires training data planned by artists and cannot control emotional types. These methods all require driving fixed facial templates in the dataset to generate speaking faces, and cannot generate any customized speaking facial styles; Moreover, the expressive power of the faces generated by these methods is limited, with a single and rigid expression. In contrast, our method can generate faces of any specified style from semantics, and can control emotional types.

2.4 Transformers and Stable Diffusion Networks

The advent of the attention mechanism is rooted in the nuanced study of human vision and has evolved as a cornerstone in an array of applications spanning natural language processing, statistical learning, image detection, and speech recognition, particularly with the exponential progression of deep learning technologies. Fundamentally, the attention mechanism is designed for the optimal allocation of informational processing resources. It mimics the human tendency to prioritize focal points within a scene while relegating peripheral, less critical elements to the background. Through assigning higher weights to pivotal

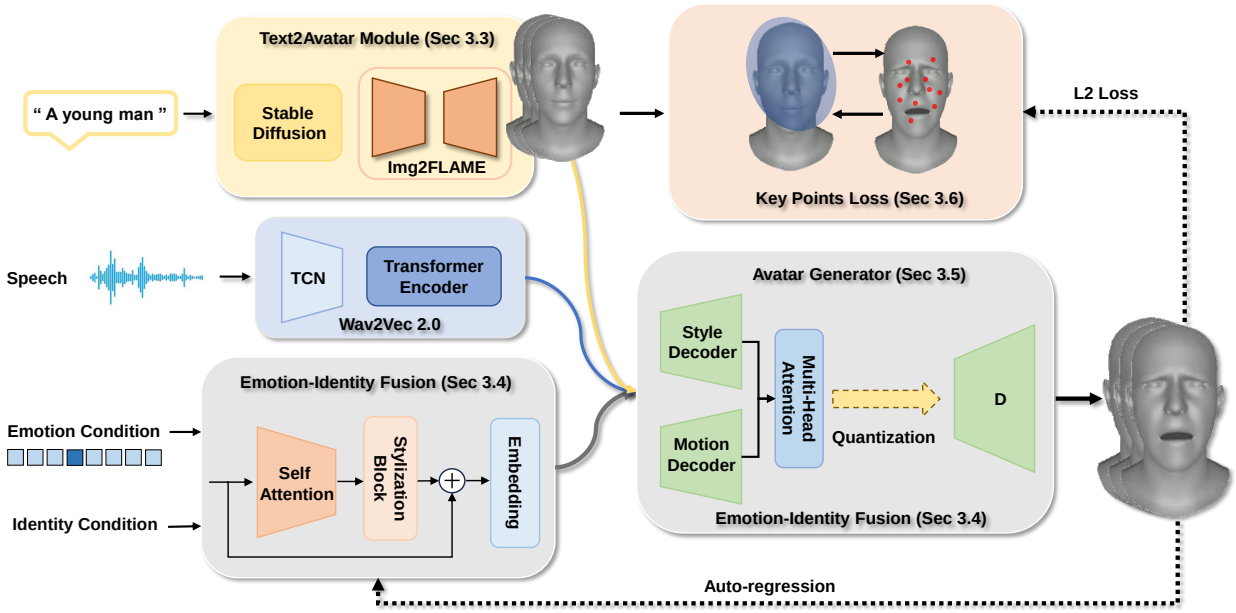


Fig. 2: The framework of our method. Input speech audio, specified facial expression types, and prompt words for customized facial templates, and then generate a 3D talking facial model.

information and lower weights to less relevant data, the attention mechanism stands distinguished by its efficiency, speed, and performance, outpacing conventional CNNs and RNNs.

Transition to Transformer Architecture The transformative impact of the Transformer model [40], initially unveiled in the domain of NLP, lies in its adept incorporation of the self-attention mechanism. This model excels in encoding and representing input sequences, capturing intricate dependency relationships amongst diverse sequence positions to foster context-awareness. The Vision Transformer (ViT) [16], a derivative adaptation for image classification, exemplifies the versatility and efficacy of Transformer models. ViTs have transcended their initial applications, permeating into various computer vision domains including object detection and semantic segmentation.

In the context of 3D facial animations, Transformer models have demonstrated a prolific capacity to generate animations that are both realistic and expressive, derived from nuanced audio descriptions. Our methodology is anchored in this architecture, integrating the Transformer's innovative design to enhance the expressive capacity of our models.

Integration of Stable Diffusion Stable diffusion, originating from Latent Diffusion Models (LDMs) [34], emerges as a potent text-to-image generation model. With the computational backing of Stability AI and data support from LAION, a specialized LDM has been trained on a subset of LAION-5B, marking a significant stride in text and image generation. The model, characterized by its iterative "denoising" process in potential representation spaces and subsequent decoding into comprehensive images, has made text and image generation accessible on consumer-grade GPUs.

In the scope of our research, stable diffusion serves as a pivotal mechanism for generating customized facial templates derived from textual descriptions. This integration not only amplifies the diversity of generated faces but also underscores our commitment to harnessing cutting-edge technologies to enhance the realism and expressiveness of 3D facial animations.

3 Method

3.1 Overview

In order to obtain customizable character appearances and emotionally rich facial speech motions, we propose a new method aimed at achieving the conversion from speech to 3D facial motions under specified speaker identities and emotions. Previous works are able to produce accurate lip movements, but the character's exterior style is rigid and cannot be customized. Therefore, we expect to obtain facial animations with customizable appearance styles. The specific pipeline of our network is shown in Fig. 2. The entire network consists of four modules. Firstly, pre-trained Wav2Vec2.0 [4] model are used to encode audio to obtain voice signals. Secondly, the Text2Avatar Module receives customized prompts for the appearance of the speaker, thereby generating the desired facial template. Thirdly, the Emotion-Identity Fusion Module embeds the specified expression into facial movements. Finally, the Avatar Generator uses the encoded signal from the previous module to generate facial animation through an auto-regressive process, querying facial motion priors within the constrained space of the codebook. The entire goal can be described as follows:

$$\hat{\mathbf{M}} = F(\mathbf{A}, \mathbf{Temp}, \mathbf{E}, \mathbf{I}), \tag{1}$$

where $\mathbf{A}$ represents the driving audio, $\mathbf{Temp}$ represents the customized facial appearance prompt texts, $\mathbf{E}$ represents the conditioned emotions, $\mathbf{I}$ represents the talking style, F is our method and $\hat{\mathbf{M}}$ is the predicted facial motion. By leveraging these inputs, our method is capable of generating 3D talking faces with diverse styles in response to the provided audio.

3.2 Preliminaries

Facial motion is a complex phenomenon, as a single lip shape can correspond to multiple facial expressions, and a single speech can be accompanied by various potential facial expressions. It is advantageous to map high-dimensional complex faces into low-dimensional spaces to simplify the mapping process and alleviate ambiguity (especially in one-to-many cases). Additionally, it is important to note that facial movements are interconnected and not independent. Consequently, our approach involves training autoencoders to learn motion sequences as a preliminary step.

Avatar Facial Motion Primitives. Given the facial motion $\mathbf{M}_{1:T}$, the encoder maps it to a continuous latent, this process can be represented as $\mathbf{Z_e} = \mathrm{E}(\mathbf{M}_{1:T})$. Then a vector quantization function $Q(\cdot)$ map to a discrete latent representation $\mathbf{Z_q}$, this process can be formulated as follows:

$$\mathbf{Z_q} = Q(\mathbf{Z_e}) = \mathrm{argmin}_{e_i} \|\mathbf{z_e} - \mathbf{e_i}\|_2^2. \tag{2}$$

Then, the self-reconstruction process can be expressed as follows:

$$\hat{\mathbf{M}}_{1:T} = D\left(\mathbf{Z_q}\right) = D\left(Q\left(E\left(\mathbf{M}_{1:T}\right)\right)\right), \tag{3}$$

where $\hat{\mathbf{M}}_{1:T}$ is the generated facial action, and $D(\cdot)$ is the motion decoder.

Facial Losses design. In order to better train the quantized autoencoder, we have carefully designed the following loss functions to achieve the best results:

$$L = L_{\text{rec}} + \beta L_{\text{vq}}, \tag{4}$$

$$L_{\text{rec}} = \|\mathbf{M}_{1:T} - \hat{\mathbf{M}}_{1:T}\|_1, \tag{5}$$

$$L_{\text{vq}} = \|\,\text{sg}(\mathbf{Z_e}) - \mathbf{Z_q}\|_2^2 + \|\mathbf{Z_e} - \text{sg}(\mathbf{Z_q})\|_2^2, \tag{6}$$

Our loss mainly consists of two parts, namely reconstruction loss and vq loss, and we have added a coefficient β to balance the two. The first term is a reconstruction loss. It involves calculating the mean absolute error between the ground truth motions and the predicted motions for each frame $\mathbf{m}_t$ in the sequence. The following two components are intermediate losses used to update the codebook items. This is achieved by minimizing the distance between the codebook vectors $\mathbf{Zq}$ and the embedded features $\mathbf{Z_e}$. The stop-gradient operation $\text{sg}(\cdot)$ is employed as a crucial mechanism to prevent collapsing.

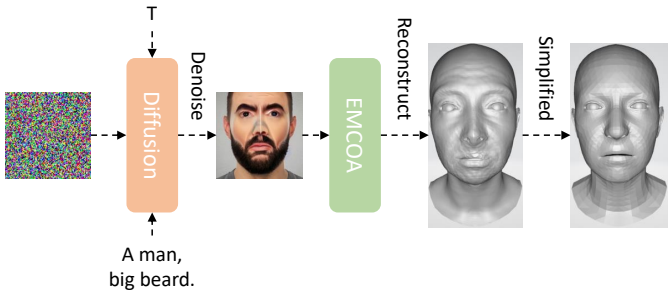


Fig. 3: Display of reconstructed facial images. The input on the left is the facial image generated by the stable diffusion, the output of the EMCOA model in the middle, and the simplified result based on the Flame model on the right.

3.3 Text2Avatar

To create customized character appearances, we design a Text2Avatar Module. This module takes character prompts as input and generates custom facial templates. These templates are then used by the Avatar Generator module to produce facial speech animations with different emotions. The input for this process consists of several text keywords, and the output is a static human head model.

The workflow of the Text2Avatar module begins by inputting the prompt words. Firstly, we use a stable diffusion mode [33] to generate a facial image. This defines a Markov chain of diffusion steps, slowly adding random noise to the real data, called a forward process. Then, we learn the reverse diffusion process, called a reverse process, which constructs the image data we need from the noise using conditional terms. This image is fed into the facial reconstruction model EMOCA [13], producing the head template. From the reconstructed head template, we select 5023 vertices based on the Flame model to create a simplified model, resulting in the best performance of the generated facial model while preserving maximum detail. This process can be formulated as follows:

$$Temp = Simp(EMOCA(Diff(C, ST))), \tag{7}$$

where C is the facial prompt word, ST is the number of steps sampled by the diffusion model, $Diff$ represents the diffusion model, $EMOCA$ is the facial reconstruction model, $Simp$ represents the simplification of the facial template, and $Temp$ is the final generated facial template.

The specific implementation is shown in Fig. 3, the reconstructed model from the image effectively restores the details of the face, while the simplified model retains most of the original facial style.

3.4 Emotion-Identity Fusion Module

The emotion identity fusion module achieves facial expression control and can learn character language style clues. In the module, in order to better embed emotion and identity vectors into facial motions, Stylization Block [47] is applied after Linear Self attention [47], which brings timestamp t to the generation process. Specifically, two features identity i and emotion e are fused with Linear Self-attention, outputs pass through Stylization Block and are added to the information. The embedding block combines the past facial motions $M_{1:t-1}$ and the style embedding via:

$$\mathbf{F}_{emb}^{1:t-1} = L_\theta(\hat{\mathbf{M}}_{1:t-1}) + S(Attn(\mathbf{B}\cdot\frac{\mathbf{i}}{||\mathbf{i}||_1}, \mathbf{B}\cdot\frac{\mathbf{e}}{||\mathbf{e}||_1}), \mathbf{B}\cdot\frac{\mathbf{e}}{||\mathbf{e}||_1}), \tag{8}$$

where $\mathbf{F}_{emb}^{1:t-1}$ represents the encoded feature, $L(\cdot)$ is the projection layer, $Attn(\cdot)$ is the Linear Self-attention, $S(\cdot)$ is the Stylization Block, $\mathbf{B}$ represents the adaptable fundamental vectors that form a linear span across the emotion and identity domain.

The advantage of Linear Self-attention is that feature map focus more on global information while classical self-attention focus more on pairwise relation. The semantic clues provided by global information are beneficial for the coordination of facial motion synthesis. Embedding style features through linear transformation, and obtaining timestamp embedding through positional embedding. The output of the attention mechanism is added to the information through the Stylization Block, and the module incorporates timestamp t and rich expression semantics into the generation process.

3.5 Avatar Generator

We develop an Avatar Generator based on the attention mechanism to generate talking avatars with different appearances, styles, and corresponding speech. Firstly, we employ a self-supervised pretrained speech model Wav2Vec 2.0 as a primary component to encode the speech. This model comprises an audio feature extractor and a multilayer transformer encoder. The audio feature extractor transforms raw waveform speech into feature vectors using a temporal convolutional network (TCN). Subsequently, the transformer encoder contextualizes these feature vectors, resulting in contextualized speech features:

$$\mathbf{F}_s^{1:T} = W(TCN(\mathbf{A})), \tag{9}$$

where $\mathbf{F}_s^{1:T}$ represents the audio features of the encoded frames 1 to T, and W represents the Transformer Encoder.

Next, we use two transformer decoders equipped with causal self-attention to learn the correlation between each frame in the context of the facial motion sequences. Specifically, the Style Decoder focuses on decoding style and emotional features, while the Motion Decoder focuses on facial motion details. The outputs of the two decoders are fused together with multi-head attention, aligning the audio and motion modalities and preserving rich facial expressions. this process can be formulated as follows:

$$\hat{\mathbf{Z}}^{1:t} = MultiAttn(D_m(L_\theta(\hat{\mathbf{M}}_{1:t-1}), \mathbf{F}_s^{1:T}), D_s(\mathbf{F}_{emb}^{1:t-1}, \mathbf{F}_s^{1:T})), \tag{10}$$

where $\hat{\mathbf{Z}}^{1:t}$ is the output features, D_m is the motion decoder, D_s is the style ecoder.$\hat{\mathbf{Z}}^{1:t}$ is further quantified into $\mathbf{Z}^{1:t}$ via Eq. 2 and decoded by the first stage pre-trained VQ-VAE decoder $\mathbf{D}$, ultimately generating $\mathbf{M}_{1:t}$.

3.6 Key Points Loss

Our loss function comprises four distinct components: regression loss, motion loss and key points loss. The overall function is given by:

$$L_{\text{syn}} = \lambda_1 * L_{\text{reg}} + \lambda_2 * L_{\text{motion}} + \lambda_3 * L_{\text{key}} \tag{11}$$

$$L_{\text{reg}} = \left\|\hat{\mathbf{Z}}^{1:T} - \text{sg}\left(\mathbf{Z}_{\mathbf{q}}^{1:T}\right)\right\|_2^2, \tag{12}$$

$$L_{\text{motion}} = \left\|\hat{\mathbf{M}}_{1:T} - \mathbf{M}_{1:T}\right\|_2^2, \tag{13}$$

$$L_{\text{key}} = \left\|\hat{\mathbf{M}}^k_{1:T} - \mathbf{M}^k_{1:T}\right\|_2^2. \tag{14}$$

The regularization loss L_{reg} acts as a guide for the model's predictions $\hat{\mathbf{Z}}^{1:t}$, encouraging them to align more closely with the quantized values $\mathbf{Z}^{1:t}$ and promoting regularity in the predicted motion sequences. On the other hand, the motion loss L_{motion} measures the disparity between predicted motions and actual motions, thereby enhancing the authenticity of facial expressions. Additionally, we introduce L_{key}, which emphasizes the difference between predicted and actual values of "active key points". This approach injects dynamism into rigid facial expressions and improves lip accuracy. The operational principle is as follows: Firstly, random facial motion sequences with diverse expressions are selected from the dataset, and then the variance of distance values for each facial registration vertex across all frames is calculated. This variance, denoted as $var = \sum_{j=1}^{V} \frac{1}{N} \sum_{i=1}^{N} (d_{ij} - d_i)^2 \in R^V$, highlights significant changes in distance between different frames and the natural expression state of the template object. Subsequently, the top k vertices with the highest variances are chosen to form the set of active key points, denoted as $S = Top_k(var) \in R^k$. Finally, calculate the L2 error between the predicted and the ground truth of these k points. This approach allows for active movement of the facial muscles, enabling key parts of the face to better convey emotions.

4 Experiments

4.1 Experimental Settings

Datasets: We selected two people's videos from the MEAD [42] dataset and reconstructed them into a 3D flame mesh as the training model dataset. Each subject has a total of 8 expressions, and we only select videos with level 3 expression intensity. Each expression has approximately 30 sentences. Our dataset has a total audio duration of approximately 0.58 hours, which is 0.48 for the VOCASET [11] dataset and 1.43 for the BIWI [18] dataset.

We utlize state-of-the-art face reconstruction methods EMOCA [12] to reconstruct high-detailed 3D facial expressions from single monocular images to FLAME [24] model. The 3D faces sequences are captured at 25fps, each mesh is registered to the FLAME topology with 5023 vertices and one subject has one template. The audio is converted to 16000Hz and is clipped into segments using a sliding window. Wav2Vec2.0 [4] uses the model provided by the original paper to extract audio features.

Implementation Details: Motion prior is trained on the reconstructed dataset for 200 epochs with batch size 1. AdamW is used as optimizer, with a learning rate of 1e-4. The size of the latent space is set empirically to 128 and q = 8. For training animation model, we train it with Adam optimizer with a learning rate of 1e-5 for 500 epochs.

4.2 Evaluation Methods

Lip Movement Error(LME): For a fair comparison in lip synchronization, we used the FLAME parameters to compare with other prior works quantitatively. We select 254 vertices around the lips and calculate the maximum L2 error from the ground truth as measurements of lip synchronization.

Active Points Error(AVE): We also compared the errors of active key points. Using the highest variance method introduced in the previous section on the Flame model, select 1000 points and compare their maximum L2 error, which can compare whether the facial expression ability is active or rigid.

4.3 Quantitative Evaluation

Comparison with SOTA methods We selected two individuals from the dataset after facial reconstruction that were different from the training set as the test subjects, and selected all 8 expressions with 5 sentences for each expression. Table 1 shows that our method achieved the best results compared to all other methods, indicating that our proposed process can generate good lip shape while enhancing emotions.

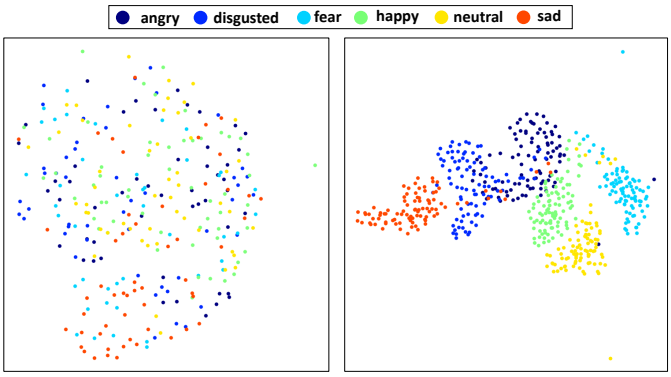
Table 1: Quantitative evaluation on reconstructed dataset. Lower means better fort both metrics.

Method	LME ↓	AVE ↓
FaceFormer [17]	1.2071	1.2857
CodeTalker [46]	1.3039	1.2637
Ours	**1.2070**	**1.2528**

Ablation experiments Table 2 shows the effect of our proposed key active point loss on lip synchronization. We observed that the proposed method contributes to the accuracy of lip movement, and deletion can lead to an increase in erroneous indicators.

Table 2: Ablation evaluation on reconstructed dataset.

Method	LME ↓	AVE ↓
w/o lossmouth	1.2931	1.3604
w lossmouth	1.2070	1.2528



(a) w/o emotion-identity fusion module (b) w/ emotion-identity fusion module

Fig. 4: The t-SNE diagram of the same sentence under different emotions types.

Fig. 4 shows the effectiveness of the emotion and identity fusion module. We use t-SNE to visualize results of the same sentence under different emotional styles of the same character. From the figure, it can be seen that the expressions before using the module are difficult to distinguish. After using the module, different expressions are roughly divided into different regions, and the expressions are more distinct. This proves that our proposed method can achieve good visual effects.

4.4 Qualitative Evaluation

We plotted the changes in the distance between the upper and lower lips to visually display the lip movement results of different models in the Fig. 10. As shown in Figure, in terms of lip changes, we observe that CodeTalker is slightly better than FaceFormer, and our method has a much larger change amplitude than FaceFormer. Overall, our method has a greater degree than both. This indicates that our method generates more expressive facial animations while ensuring lip accuracy.

Emotion control. Fig. 7 shows that our method can generate a specified facial template based on prompt words, and then generate 3D speaking faces with different expressions. The prompt word in the first line of the image is "a man", and the prompt word in the second image is "a girl".

Lip synchronization. Fig. 5 shows the lip synchronization performance by selecting frames corresponding to specific syllables in the synthesized facial animation. The results indicate that our model can generate lip movements that accurately express speech signals under different facial expressions. We also displayed the frames of key syllables in the entire sentence, as shown in Fig. 6, which can obtain accurate mouth movements.

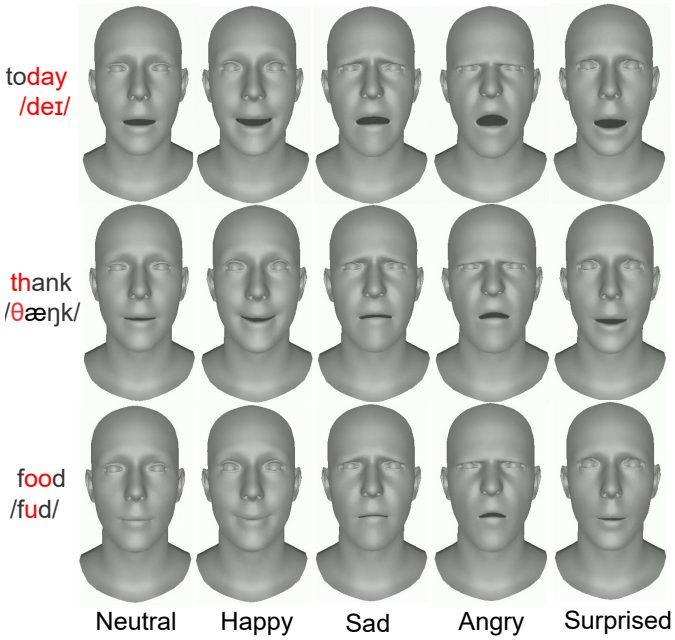


Fig. 5: We compared the mouth movements of our method on different expressions on the same syllable. Show that our method can exhibit precise lip movements under different facial expressions.

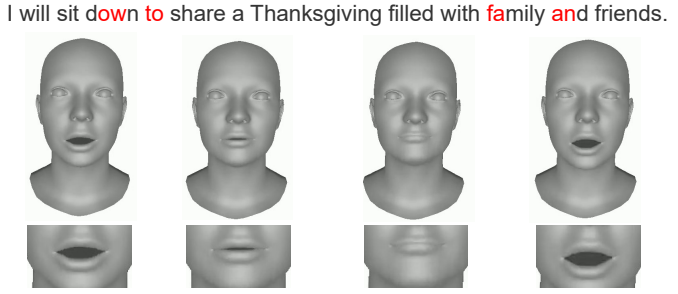


Fig. 6: Visualization of facial movements for different statements. The image sequence corresponds to the red syllable sequence in the text. The output of our method has a good visual perception.

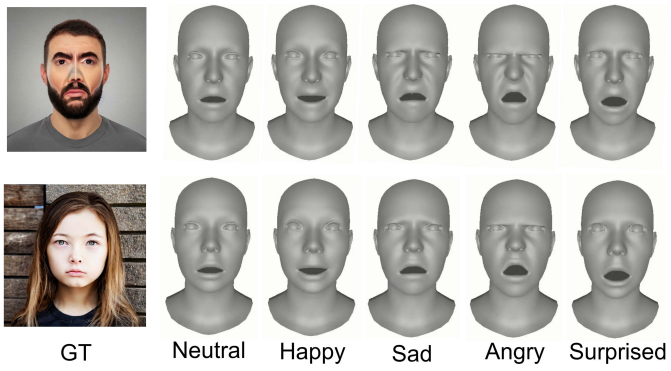


Fig. 7: Cross modal generation of talking faces with different emotions.

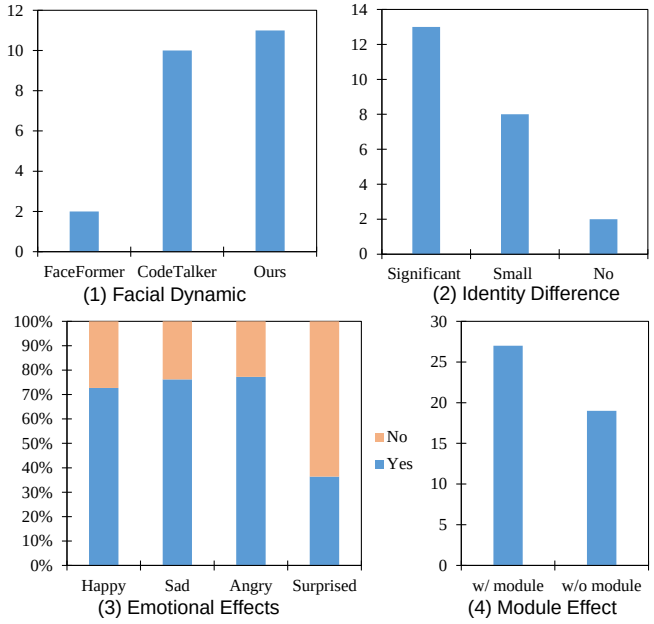


Fig. 8: We present different user research results from (1) to (4) in the figure.

Ablation experiments. Fig. 9 shows the effect of the emotion-identity fusion module. As shown in the red box, on the muscles of the eyes and cheeks, those with modules can produce more obvious and intense movements than those without them, which is more expressive and can convey more emotional clues.

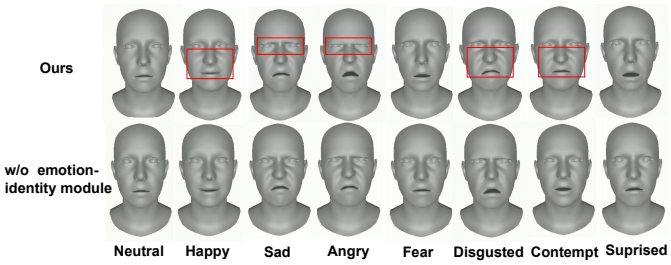


Fig. 9: Ablation experiment on the quality of emotion-identity fusion module.

4.5 User Study

Humans can understand subtle facial movements, therefore, it is a reliable metric in speech driven facial animation tasks. We conducted a user study to evaluate the facial dynamics of animated faces compared to the SOTA methods FaceFormer and CodeTalker. In addition, to verify the effectiveness of our proposed method, we conducted ablation comparison experiments on lip synchronization and emotional expression realism.

We published a questionnaire online for evaluation and ultimately received a total of 22 results. The leading age group of recruiters is between 20 and 30 (21 individuals), with the remaining age groups distributed between 30 and 50. Regarding occupational distribution, 21 are college students, and 1 is a teacher, We design a total of 6 comparisons, each comprising 5 samples that compare our method with other methods. Participants are asked to choose the best-performing one from several videos, and the final result is the proportion of our method to the total.

The quantitative indicators of all studies are presented in Table 3. In order to compare the differences between different methods, we have also drawn bar charts and scale charts to demonstrate the superiority

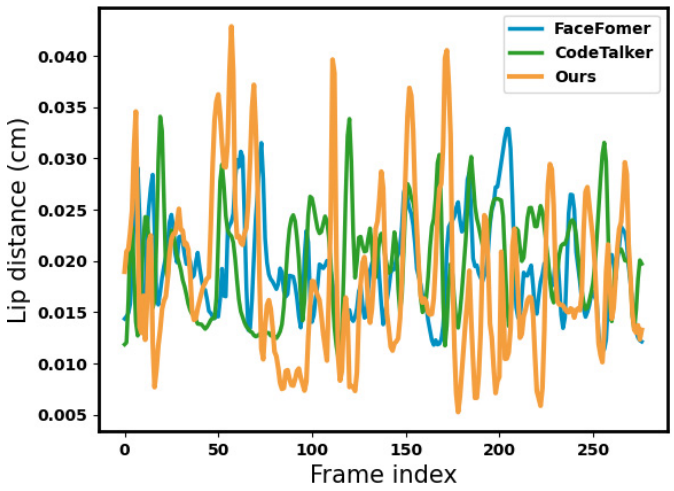


Fig. 10: Visualization of lip distance in conversation sequences of different models.

Methods	Facial dynamism	Sync quality	Emotion expression
Ours vs. FaceFormer	**84.6%**	-	-
Ours vs. Codetalker	**52.4%**	-	-
Ours vs. w/o L_{key}	**63.6%**	**52.2%**	-
Ours vs. w/o e-i fusion module	-	-	**58.7%**

Table 3: Perceptual study results on facial dynamics, lip synchronization, and the effectiveness of our proposed method.

of our method, and the results are shown in Fig. 8. From these results, we can see that the subjects are more inclined to use our method. In terms of facial dynamics, our method can achieve a significant advantage of 84.6% compared to Faceformer, and 52.4% compared to CodeTalker. In order to verify the L_{key} we proposed, we conducted ablation experiments, and the results showed that we achieved 63.6% and 52.2% advantages in facial dynamics and lip synchronization quality, respectively. In terms of facial expression, our method with the emotion-identity fusion module has a 58.7% advantage. We also demonstrate the ability of the model to specify speaker identity in Fig. 8 (2). We change the identity of different characters to speak the same paragraph while ensuring other conditions are the same. 91.3% of the respondents believe that they can more or less distinguish the style of speech.

5 Limations And Future Work

Compared to traditional speech facial generation methods that focus too much on lip accuracy, although our method can customize identity features and generate controllable expressions, the expression styles are not rich enough and the degree of controllability is still limited, and there is still a lack of facial realism. Our main goals for future research include further enriching the small details and expression styles of facial movements during speech, striving to achieve greater controllability, making them more realistic and lifelike. In addition, we will explore the potential of speaking faces and find new modes suitable for VR scenarios.

6 Conclusion

Our study focused on making 3D facial animations more lifelike by adding real emotions and unique characteristics to each face. Until now, a lot of focus has been on making the lips move realistically with speech, but the animations lacked the warmth of emotions and the uniqueness of individual identities. We introduced a new method that uses both speech and user-provided words to create 3D faces that not only talk but also express feelings and have their own identity. We used advanced techniques to make sure every animation tells a story, not just through words spoken but also through the emotions expressed and the uniqueness of the character. In simpler terms, our work is about giving life to 3D animations. We have moved beyond just lip movement and speech to create animations that feel real, have their own identity, and can express emotions, making each one special and relatable. This is a big step forward in the world of 3D facial animations. Our method still has limitations in generating the facial animation. The main limitation is that our avatar's shape is limited by the FLAME template which lacks the geometry details. In our future work, we would extend the avatar's quality in terms of geometry details and textures.

References

[1] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8717–8727, 2018. 2

[2] T. Afouras, J. S. Chung, and A. Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018. 2

[3] H. Averbuch-Elor, D. Cohen-Or, J. Kopf, and M. F. Cohen. Bringing portraits to life. *ACM transactions on graphics (TOG)*, 36(6):1–13, 2017. 2

[4] A. Baevski, H. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv e-prints*, p. arXiv:2006.11477, June 2020. doi: 10.48550/arXiv.2006.11477 3, 5

[5] L. Chen, R. K. Maddox, Z. Duan, and C. Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7832–7841, 2019. 2

[6] Y. Chen, J. Zhao, and W. Zhang. Expressive speech-driven facial animation with controllable emotions. *CoRR*, abs/2301.02008, 2023. doi: 10.48550/arXiv.2301.02008 2

[7] K. Cheng, X. Cun, Y. Zhang, M. Xia, F. Yin, M. Zhu, X. Wang, J. Wang, and N. Wang. Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–9, 2022. 2

[8] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8188–8197, 2020. 2

[9] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 2

[10] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. Black. Capture, learning, and synthesis of 3D speaking styles. *Computer Vision and Pattern Recognition (CVPR)*, pp. 10101–10111, 2019. 2

[11] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10101–10111, 2019. 1, 2, 5

[12] R. Danecek, M. J. Black, and T. Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20311–20322, 2022. 1, 5

[13] R. Daněček, M. J. Black, and T. Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20311–20322, 2022. 4

[14] S. d'Apolito, D. P. Paudel, Z. Huang, A. Romero, and L. Van Gool. Ganmut: Learning interpretable conditional space for gamut of emotions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 568–577, 2021. 2

[15] H. Ding, K. Sricharan, and R. Chellappa. Exprgan: Facial expression editing with controllable expression intensity. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018. 2

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, and T. Unterthiner. Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[17] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18770–18780, 2022. 1, 2, 5

[18] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. V. Gool. A 3-d audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 12(6):591 – 598, October 2010. 1, 5

[19] H. Fang, D. Weng, Z. Tian, and Z. Song. Audio to deep visual: Speaking mouth generation based on 3d sparse landmarks. In *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops, VR Workshops 2023, Shanghai, China, March 25-29, 2023*, pp. 605–606. IEEE, 2023. doi: 10.1109/VRW58643.2023.00145 2

[20] X. Ji, H. Zhou, K. Wang, W. Wu, C. C. Loy, X. Cao, and F. Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14080–14089, 2021. 2

[21] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 1, 2

[22] R. Kato, Y. Kikuchi, V. Yem, and Y. Ikei. Cv-mora based lip sync facial animations for japanese speech. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops, VR Workshops, Christchurch, New Zealand, March 12-16, 2022*, pp. 558–559. IEEE, 2022. doi: 10.1109/VRW55335.2022.00130 1

[23] H. Kim, M. Elgharib, M. Zollhöfer, H.-P. Seidel, T. Beeler, C. Richardt, and C. Theobalt. Neural style-preserving visual dubbing. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019. 2

[24] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 2, 5

[25] A. Lindt, P. Barros, H. Siqueira, and S. Wermter. Facial expression editing with continuous emotion labels. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1–8. IEEE, 2019. 2

[26] B. Liu, X. Wei, B. Li, J. Cao, and Y. Lai. Pose-controllable 3d facial animation synthesis using hierarchical audio-vertex attention. *CoRR*, abs/2302.12532, 2023. doi: 10.48550/arXiv.2302.12532 2

[27] R. Nishimura, N. Sakata, K. Harada, T. Tominaga, K. Kiyokawa, and Y. Hijikata. Speech-driven facial animation by LSTM-RNN for communication use. In *IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2019, Osaka, Japan, March 23-27, 2019*, pp. 1102–1103. IEEE, 2019. doi: 10.1109/VR.2019.8798145 1

[28] Y. Pan, R. Zhang, J. Wang, N. Chen, Y. Qiu, Y. Ding, and K. Mitchell. Miencap: Performance-based facial animation with live mood dynamics. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops, VR Workshops, Christchurch, New Zealand, March 12-16, 2022*, pp. 654–655. IEEE, 2022. doi: 10.1109/VRW55335.2022.00178 1

[29] Z. Peng, H. Wu, Z. Song, H. Xu, X. Zhu, J. He, H. Liu, and Z. Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. 2023. 2

[30] H. Pham, S. Cheung, and V. Pavlovic. Speech-driven 3d facial animation with implicit emotional awareness: A deep learning approach. In *IEEE Int'l Conf. Computer Vision and Pattern Recognition - Workshop on Deep Affective Learning and Context Modeling*, 2017. doi: 10.1109/cvprw.2017.287 1, 2

[31] H. X. Pham, Y. Wang, and V. Pavlovic. End-to-end learning for 3d facial animation from raw waveforms of speech. *arXiv preprint arXiv:1710.00920*, 2017. 1, 2

[32] A. Richard, M. Zollhöfer, Y. Wen, F. De la Torre, and Y. Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1173–1182, 2021. 1, 2

[33] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2021. 4

[34] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022. 3

[35] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018. 2

[36] B. Thambiraja, I. Habibie, S. Aliakbarian, D. Cosker, C. Theobalt, and J. Thies. Imitator: Personalized speech-driven 3d facial animation. *CoRR*, abs/2301.00023, 2023. doi: 10.48550/arXiv.2301.00023 2

[37] B. Thambiraja, I. Habibie, S. Aliakbarian, D. Cosker, C. Theobalt, and J. Thies. Imitator: Personalized speech-driven 3d facial animation, 2023. doi: 10.48550/ARXIV.2301.00023 2

[38] S. Tripathy, J. Kannala, and E. Rahtu. Facegan: Facial attribute controllable reenactment gan. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1329–1338, 2021. 2

[39] A. van den Oord, O. Vinyals, and k. kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017. 2

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3

[41] M. Volonte, E. Ofek, K. Jakubzak, S. Bruner, and M. González-Franco. Headbox: A facial blendshape animation toolkit for the microsoft rocketbox library. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops, VR Workshops, Christchurch, New Zealand, March 12-16, 2022*, pp. 39–42. IEEE, 2022. doi: 10.1109/VRW55335.2022.00015 1

[42] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy. MEAD: A large-scale audio-visual dataset for emotional talking-face generation. pp. 700–717, 2020. 1, 2, 5

[43] X. Wang, Y. Wang, W. Li, Z. Du, and D. Huang. Facial expression animation by landmark guided residual module. *IEEE Trans. Affect. Comput.*, 14(2):878–894, 2023. doi: 10.1109/TAFFC.2021.3100352 2

[44] R. Wu, Y. Yu, F. Zhan, J. Zhang, X. Zhang, and S. Lu. Audio-driven talking face generation with diverse yet realistic facial animations. *Pattern Recognit.*, 144:109865, 2023. doi: 10.1016/j.patcog.2023.109865 2

[45] C.-h. Wuu, N. Zheng, S. Ardisson, R. Bali, D. Belko, E. Brockmeyer, L. Evans, T. Godisart, H. Ha, X. Huang, et al. Multiface: A dataset for neural face rendering. *arXiv preprint arXiv:2207.11243*, 2022. 2

[46] J. Xing, M. Xia, Y. Zhang, X. Cun, J. Wang, and T.-T. Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. *arXiv preprint arXiv:2301.02379*, 2023. 1, 2, 5

[47] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 4

[48] P. Zhou, Y. Bao, and Y. Qi. Learning detailed 3d face via CLIP model from monocular image. In *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops, VR Workshops 2023, Shanghai, China, March 25-29, 2023*, pp. 771–772. IEEE, 2023. doi: 10.1109/VRW58643.2023.00228 2

[49] H. Zhu, W. Wu, W. Zhu, L. Jiang, S. Tang, L. Zhang, Z. Liu, and C. C. Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pp. 650–667. Springer, 2022. 2