EmoTalker: Audio Driven Emotion Aware Talking Head Generation

Xiaoqian Shen, Faizan Farooq Khan, and Mohamed Elhoseiny

King Abdullah University of Science and Technology (KAUST)
{xiaoqian.shen,faizan.khan,mohamed.elhoseiny}@kaust.edu.sa

Abstract. Talking head synthesis aims to create videos of a person speaking with accurately synchronized lip movements and natural facial expressions that correspond to the driving audio. However, previous approaches have used reference frames or extra labels to control emotions and facial expressions, which disentangle utterance and expression and ignore the impact of audio fluctuations on face motions, e.g., head pose, facial expressions and emotions. In this work, we present **EmoTalker**, which generates arbitrary identities with diverse and natural facial expressions from audio, without relying on driving frames or emotion labels as input. To achieve this, we present frames as a sequence of 3D motion coefficients of 3DMM representation and separate them into lip-related coefficients and the remaining (head pose, expressions) as facial motions. To model lip movement, we start with a pre-trained audio encoder and map it to the corresponding lip representation. While for facial motions, we employ a two-stage training strategy: 1) We first project facial motions into a finite space of the codebook embedded with emotion-aware facial expression priors. 2) Moreover, a cross-modal Transformer is devised to explicitly model the correlations between audio and different types of facial motions. Experimental results and user studies show our model achieves state-of-the-art performance on the emotional audio-visual dataset and produces more realistic talking head videos with synchronized lip movement and vivid facial expressions. Our codes are available at https://github.com/xiaoqian-shen/EmoTalker.

Keywords: Talking Head Generation · Video Synthesis

1 Introduction

Talking head synthesis driven by audio, which involves generating a video of an individual with accurately synchronized lip movements and matched facial expressions, has been an active research area for several years. This task is particularly challenging due to the diversity of facial motions required to create a convincing video and the need for photo-realistic output. In addition, the potential applications of audio-driven talking head synthesis are vast, including real video editing, digital human creation, etc.

Previous works mainly focus on generating lip motion [4, 5, 16, 27] since it has a strong correlation with speech. Recent works [2, 36, 37, 46, 47] also aim

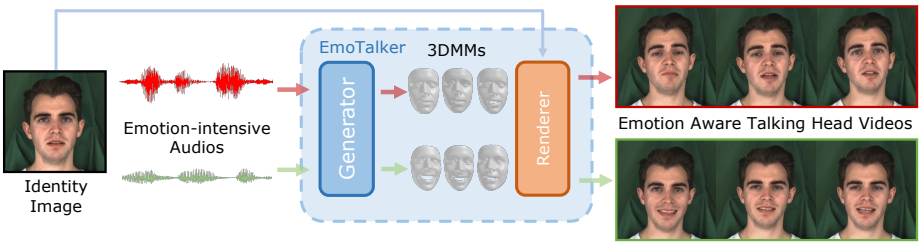

Fig. 1: Given an arbitrary identity image and emotion-rich audio as input, our model can generate talking-head videos of the identity by synchronizing lip movement with speech content and matching facial expressions with audio cadence, without using extra emotion labels or driving frames.

to generate a realistic talking face video that incorporates related motions such as head pose and facial expression. However, these approaches utilize external reference frames or emotion labels for control, which may lead to the separation of utterance and expression, disregarding the impact of audio emotions on the resulting facial expressions.

In this work, we propose a novel approach that considers the correlation between facial expressions and audio fluctuations without relying on emotion labels or predetermined frames as input control. This is because emotion is a continuously changing signal that should not be constrained by arbitrary emotion labels in real-world applications. In addition, the emotional intensity and facial expression are naturally conveyed by speech cadence. Therefore, it is more appropriate to explicitly model the correlations between audio and various types of facial motions.

Directly mapping audio signals to synchronized talking face videos is inefficient due to its high computational demands and substantial redundant information in the raw frames. Therefore, we employ 3D Morphable Models (3DMMs) as an intermediate representation for video frames, incorporating facial geometry information, facilitating the expressive manifestation of emotional facial movements. Moreover, it disentangles identity and other redundant information from the raw video, thus reducing model complexity and enabling arbitrary-identity talking video synthesis.

Although several approaches have been made to map audio to latent representations such as facial landmarks [47] or 3DMM [44], they still struggle to produce vivid facial expressions and capture the emotional signals conveyed by the audio accurately. We argue the difficulties arise from facial expressions' intrinsic diversity and flexibility, coupled with the rapid changes and precise synchronization with audio required by lip movements. Consequently, we propose a separate modeling approach for facial expressions and lip movements based on the feasibility of identifying lip-related indices of 3DMM face expression coefficients [22]. For the lip branch, we leverage a pre-trained audio encoder [27] and learn a mapping function from audio features to lip coefficient space. Given facial expressions' intricate and varied nature, learning a direct mapping function from audio to

facial motions can be challenging. To address this issue, we propose a two-stage modeling approach that enables our model to learn emotion-aware facial expression priors in the first stage and then align audio features with emotion-aware codebook embeddings in the second stage via a cross-modal Transformer. Our approach eliminates the necessity of using arbitrary emotional labels as input during inference, making the generated videos exhibit more diverse and natural facial motions consistent with the driven audio.

We evaluate our method on a large-scale emotional audio-visual dataset MEAD [35] and demonstrate that our approach produces more realistic and emotionally consistent talking head videos compared to previous methods. Overall, our work presents a significant contribution to the field of talking head synthesis and paves the way for future research in this area.

Our contributions are highlighted below:

- We argue that the motion frequency of lip movement surpasses that of other facial regions. Therefore, we propose to disentangle the 3DMM coefficients into lip-related coefficients and the remaining, which represent expressions/poses., allowing us to model them separately.
- We leverage a two-stage training strategy to model the correlation between audio and facial expressions. In the first stage, we project facial motions into a finite space of codebook embedded with emotion-aware expression priors. Then, we develop a cross-modal Transformer that retrieves facial representations in response to the input audio signal.
- Our method achieves state-of-the-art performance in emotion-aware audio-driven talking head video generation and synthesizes vivid facial expressions matched with audio cadence without the need for external emotional labels or control frames.

2 Related Work

2.1 Audio-driven Talking Head Generation

Early works [4, 27, 34, 45] mainly focus on creating accurate lip movements that are synchronized with the speech content. Recently, several methods [2, 36, 37, 46, 47] try to produce more natural faces by taking facial expression into consideration, by explicitly using expression/pose [13, 19] or style frames as reference [22, 39]. Several works use audio while several person-specific methods can generate high-fidelity videos from speech audio [10, 42], but can only work on a single subject. The comparison among recent methods in audio-driven talking face generation is detailed in supplementary in terms of representation space (raw image or 3DMM [7]), whether using driven frames, etc. In this work, we aim to generate an arbitrary identity talking head video with only an identity reference image and an audio speech as input, without any driving frames.

2.2 Emotional Talking Face Generation

Emotion plays a significant role in achieving realism in animation, but few works address it in talking face generation due to the complexity of emotion dynamics. Recent work [35] has made progress by collecting an audio-visual dataset for emotional generation, upon which [14] decomposes speech into decoupled content and emotion spaces and set emotions as one-hot vectors input to achieve emotion control, while [13] resort to an emotional video as source control. However, those methods disregard the emotional information conveyed by the original audio, leading to inconsistencies between generated facial expressions and corresponding audio cadence. In this work, we propose a method for controlling facial expression by implicitly modeling emotional information from audio without the need for extra emotion labeling as input.

2.3 VQ-VAE

Vector Quantized Variational Autoencoders (VQ-VAE), which is a variant of VAE [15], is initially proposed in [33]. VQ-VAE is composed of an autoEncoder architecture, which aims at learning reconstruction with discrete representations. Recently, VQ-VAE achieves promising performance on generative tasks across different modalities, which includes: image synthesis [8,38], text-to-image generation [29,31], speech gesture generation [1,20], motion synthesis [24,40,43] etc. The success of VQ-VAE for generation might be attributed to its decoupling of learning the discrete representation and the prior. Learning a direct mapping function from audio to facial expression is challenging due to its complexity and diversity. Therefore, we leverage VQ-VAE to compress various facial expressions over time into a finite space of codebook embedded with emotion-aware priors.

3 Methodology

As depicted in Fig. 1, taking an arbitrary identity image and a sequence of speech snippets $\mathcal{A} = \{a_i\}_{i=1}^{T_a}$ as input, our model first generates a temporal sequence of 3DMM motion coefficients and then render a talking head video $\mathcal{V} = \{v_i\}_{i=1}^{T_v}$ aligned with both speech content and audio fluctuations.

3.1 Preliminary of 3DMM Representation

Operating on raw video space is computationally demanding, and disentangling identity information can be challenging. Therefore, we leverage the deep 3D reconstruction method [7] to convert the video clip into the space of the 3D Morphable Models (3DMMs) as an intermediate representation. Furthermore, this representation contains a geometric topology of the face, a more expressive depiction for conveying emotional facial expressions than raw video. With 3DMM, a coarse 3D face mesh $\mathbf{M}_i$ of face i can be represented as an affine model of facial expression and identity code:

$$\mathbf{M}_i = \overline{\mathbf{M}} + \alpha_i \mathbf{B}_{id} + \beta_i \mathbf{B}_{exp} \tag{1}$$

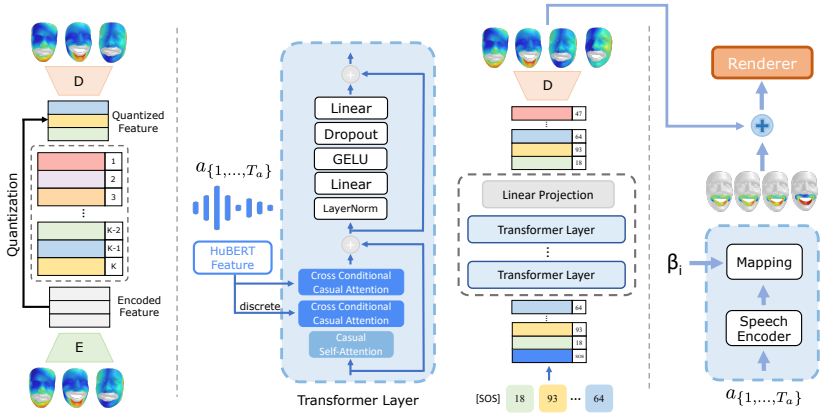

Fig. 2: The model architecture of EmoTalker. The left and middle columns are responsible for facial (non-lip) synthesis and involve two stages of training. **Left:** A VQ-VAE is trained (Sec. 3.2) to learn an emotion-enriched facial expression codebook in the first stage. **Middle:** A cross-modal Transformer decoder (Sec. 3.2) is trained to generate facial expressions that match the input audio during the second stage. **Right:** The lip-related synthesis branch (Sec. 3.2) maps raw audio $a_{\{1,\dots,T_a\}}$ to lip coefficients using reference coefficients β_i from Eq. (1). The lip-related and remaining coefficients are then combined to form the final 3DMM representation, which is fed into an off-the-shelf renderer to generate videos.

where $\overline{\mathbf{M}}$ is the average shape of the 3D face, $\mathbf{B}_{id}$ and $\mathbf{B}_{exp}$ are the bases of identity and expression computed via Principal Component Analysis (PCA) based on scans of human faces [25]; coefficients $\alpha_i \in \mathbb{R}^{80}$ and $\beta_i \in \mathbb{R}^{64}$ represent for the person identity and expression, respectively. The desired motion of face i are expressed with parameter set $p_i \equiv \{\beta_i, R_i, t_i\}$, where the head rotation and translation are expressed as $R \in SO(3)$ and $t \in \mathbb{R}^3$. To enable arbitrary identity synthesis, we disentangle identity information and use the parameter set p_i as an intermediate representation of face i during training.

3.2 Emotion Aware 3D Motion Modeling

Plainly modeling the whole face will ignore the importance of lip reconstruction quality since they are in the minority, and the frequency of motion in the lips is typically higher than other areas of the face [22]. (The variance of lip movement is 0.4925, while non-lip-related coefficients exhibit a variance of 0.1018 in the MEAD [35] dataset.) Therefore, it is more appropriate to model the lip-related and other facial coefficients separately. Inspired by previous work [22] which has identified the indices of expression parameters β highly correlated to mouth movements, we represent the facial motions of a video clip by partitioning the parameter set p_i into lip-related coefficients $m_i \in \mathbb{R}^{d_m}$ and facial coefficients $f_i \in \mathbb{R}^{d_f}$ encompassing the remaining aspects, i.e., facial expressions and head pose.

Quantized Facial Expression Codebook We leverage VQ-VAE (shown in Fig. 2 left) to recover a sequence of facial motions, which consists of an encoder E, a decoder D, and a learnable codebook denoted as $\mathcal{C} = \{c_k\}_{k=1}^{K}$ with code embedding $c_k \in \mathbb{R}^{d_c}$, where K represents the codebook size and d_c is the dimensionality of code embedding. The autoencoder includes a standard CNN-based architecture with 1D convolution, residual block and ReLU activation, using convolution with stride 2 and nearest interpolation for temporal downsampling and upsampling.

The encoded feature $Z^{1:T'_v} = [z_1, z_2, \ldots, z_{T'_v}]$ can be computed as $Z^{1:T'_v} = E(f_{1:T_v})$, where $z_i \in \mathbb{R}^{d_c}$, $T'_v = T_v/l$ and l represents the temporal downsampling rate of the encoder E. Then, we obtain the quantized motion sequence $Z_q \in \mathbb{R}^{T'_v \times d_c}$ via an element-wise quantization function $Q(.)$ that assigns each embedding in Z to its nearest entry in codebook $\mathcal{C}$:

$$Z_q = Q(Z) := \underset{c_k \in \mathcal{C}}{\arg\min} \, ||z_i - c_k|| \tag{2}$$

We add a classification head after the encoder E, which predicts the emotion labels for encoded features. After quantization, the VQ-VAE decoder is trained to reconstruct the original 3DMM feature. The training objective is formulated as below:

$$\mathcal{L}_{VQ} = \mathcal{L}_1^{smooth}(f_{1:T_v}, \hat{f}_{1:T_v}) + ||\text{sg}(Z) - Z_q||_2^2 + \beta||\text{sg}(Z_q) - Z)||_2^2 + \mathcal{L}_{cls} \tag{3}$$

where $\hat{f}_{1:T_v}$ is the reconstructed facial expression, β is a hyper-parameter for the commitment loss, sg(.) is the stop-gradient operator and $\mathcal{L}_{cls}$ is the emotion classification loss. By introducing the additional emotion classification loss $\mathcal{L}_{cls}$, the learned codebook becomes more emotion-aware, where each code represents a distinct facial expression over l frames.

Audio Driven Facial Coefficients Synthesis We utilize the HuBERT [12] model, which extracts audio embedding as a Transformer encoder. Then, we leverage the Encoder-Decoder attention as cross-modal attention in the Transformer decoder layers (see Fig. 2 middle), which conditions the audio signal to predict the temporal facial representations from the pre-trained codebook.

Biased Causal Self Attention. Attention with Linear Biases (ALiBi) [28] is proposed to improve generalization abilities for longer sequences in language modeling by adding a constant bias to the query-key attention score. Similar to [9, 40], we adopt ALiBi for facial sequence generation, which first linearly projects facial code embedding sequence, denoted as $\hat{Z}_q^{1:T'_v}$, into queries $\mathbf{Q}^{\hat{F}}$, keys $\mathbf{K}^{\hat{F}}$ of dimension d_k, and values $\mathbf{V}^{\hat{F}}$ of dimension d_v. To constrain the the dependencies on most recent frames, a weighted contextual representation is calculated by performing the scaled dot-product attention formulated as:

$$\text{SelfAtt} = \text{Softmax}\left(\frac{\mathbf{Q}^{\hat{\text{F}}}(\mathbf{K}^{\hat{\text{F}}})^{\text{T}}}{\sqrt{\text{d}_\text{k}}} + \mathbf{B}^{\hat{\text{F}}}\right)\mathbf{V}^{\hat{\text{F}}} \tag{4}$$

where

$$\mathbf{B}^{\hat{F}}_{i,j} = \begin{cases} \lfloor j - i/h \rfloor, & 1 \le j \le i \le T'_v \\ -\infty, & otherwise \end{cases} \tag{5}$$

where $\mathbf{B}^{\hat{F}}$ is a matrix that has negative infinity in the upper triangle to prevent looking ahead into future frames during prediction. Additionally, the lower triangle is designed to introduce a bias in the causal attention mechanism by assigning higher weights to the context window closer to the current frame, with context window size h.

Cross Modal Causal Attention. After the Self-Attention Layer, we proceed to compute the cross-attention between the visual and audio modalities. Despite the sample rate of video and audio aligned in raw data (i.e., 25fps vs. 16kHz), the temporal representation is modeled differently by VQ-VAE encoder (see Sec. 3.2) and HuBERT [12] audio encoder. Specifically, the HuBERT model takes an input sequence of raw audio waveforms, applies downsampling via `Conv1D`, and passes the downsampled hidden states to a Transformer network. As a consequence, a misalignment arises between the two modalities. To address this issue, we include a mask that ensures that the corresponding audio segments attend to the current facial embedding.

$$\text{CrossAtt} = \text{Softmax}\left(\frac{\mathbf{Q}^{\hat{\text{F}}}(\mathbf{K}^{\text{A}})^{\text{T}}}{\sqrt{\text{d}_\text{k}}} + \mathbf{M}\right)\mathbf{V}^{\text{A}} \tag{6}$$

where the cross-modal alignment mask $\mathbf{M}$ is defined as:

$$\mathbf{M}_{i,j} = \begin{cases} 0, & \max(1, k(i-1)) \le j \le \min(k(i+1), T'_a) \\ -\infty, & otherwise \end{cases} \tag{7}$$

where $k = \lfloor T'_a / T'_v \rfloor$, $T'_v = T_v / l$ is the temporal dimension of quantized facial expression after VQVAE encoder and the T'_a is the temporal dimension of HuBERT [12] feature. This mask restricts the attention to the specific range where the prediction of facial features only considers its corresponding audio tokens at the same time step.

Emotion-content Disentanglement. We follow [17], which involves an additional k-means step on top of the continuous representation of HuBERT outputs. This k-means step serves to quantize the continuous representations into a discrete sequence denoted as $z_c = (z_c^1, ..., z_c^{T'_a})$, where $z_c^i \in \{1, ..., K\}$, K is the size of the vocabulary, T'_a is the temporal dimension of HuBERT feature. We extracted representations from the 9-th layer of the HuBERT model and set $K = 200$. As demonstrated in [18, 26], this discrete representation effectively suppresses emotional signals while retaining content-related information. We opt to employ the cluster center embedding as the discrete code feature and add an additional cross-attention layer to interact between the output of the self-attention layer and the audio content features, i.e., the discrete HuBERT feature. Then the output of this cross-attention layer will interact with the emotion-aware audio feature.

Training Objective. Similar to [40], we allow the transformer to generate facial embeddings rather than predicting categorical distributions. The model is trained by minimizing the L_2 distance between the Transformer outputs $\hat{Z}^{1:T'_v}$ and the ground truth sequence of quantized feature $Z_q^{1:T'_v}$. To introduce diversity in facial motions and distinguish predicted facial features across different time steps, we include a contrastive loss in our approach. Considering all the features from other time steps in the video as negative samples could be problematic, as there might be instances where the same token embedding appears at different time steps. Therefore, we propose utilizing the contrastive loss to involve all token embeddings in the pre-trained codebook, except for the anchor, as negative samples. The final loss function is formulated as:

$$\mathcal{L}_{trans} = \|\hat{Z}^{1:T'_v} - Z_q^{1:T'_v}\|_2^2 + \mathcal{L}_{contra} \tag{8}$$

where

$$\mathcal{L}_{contra} = -\frac{1}{T'_v}\sum_{i=1}^{T'_v} \log \frac{e^{sim(\hat{z}_i, c_i)}}{e^{sim(\hat{z}_i, c_i)} + \sum_{j=1}^{K} e^{sim(\hat{z}_i, c_{ij}^-)}} \tag{9}$$

where $sim(,)$ computes the cosine similarity and c_i represents the code embedding from the learned codebook. We take each positive pair (z_i, c_i) and its associated set of negative pairs (z_i, c_{ij}^-), and pass them through a Softmax layer with multi-class logistic loss. This process encourages the similarity between positive pairs to approach 1, and 0 otherwise. By employing this loss, our model gains discriminative power to distinguish among different facial representations, while also naturally promoting alignment between audio snippets and their corresponding facial features.

Audio Driven Lip Coefficients Synthesis To model the high-frequency lip movement, we disentangle the lip-related coefficients from the 3DMM representation and take audio and reference expression coefficients as input to predict the lip motion. More specifically, a mapping network ϕ_{M_1} maps the concatenation of randomly sampled expression coefficients β_i, $i \in [1, T_v]$ and the audio feature extracted by pre-trained speech encoder [27] Φ_A into expression coefficients following [44]. Then an additional mapping network Φ_{M2} is proposed to predict the lip-related coefficients, formulated as follows:

$$m_{\{1,..,T_v\}} = \Phi_{M_2}(\Phi_{M_1}(\Phi_A(a_{\{1,..,T_a\}}), \beta_i)) \tag{10}$$

where $m_{\{1,..,T_v\}}$ is the predicted lip-related coefficients. Intuitively, the mapping layer Φ_{M_2} distills the knowledge from pre-trained weight and refines the mouth shape of the predicted expression coefficients to better match the predicted facial expressions.

Inference Phase During the inference phase, the pre-trained autoencoder from the first stage encodes the 3DMM representation extracted from the reference

Method	Texture Quality		Landmark Alignment		Lip Sync	Emotion
	SSIM ↑	CPBD ↑	F-LMD ↓	M-LMD ↓	Sync_{conf} ↑	Emo_{Acc} ↑
Real Video	1.	0.458	0.	0.	2.180	1.
MakeItTalk [47]	0.671	0.320	10.188	10.362	2.104	0.273
Wav2Lip* [27]	0.654	0.402	10.062	10.288	2.688	0.128
Audio2Head [36]	0.684	0.249	8.372	8.544	1.887	0.193
SadTalker [44]	0.701	0.353	4.764	5.620	2.120	0.296
EmoTalker (ours)	**0.705**	**0.362**	**4.348**	**5.139**	**2.304**	**0.425**

Table 1: Quantitative comparison on MEAD [35] dataset. To prevent any emotional bias, all the models take a neutral reference frame as input. M- represents mouth and F- stands for face region. *Wav2Lip [27] achieves the best video quality since it only animates the lip region and copy-paste the generated lip to the original frame.

image. The encoded feature, along with the input audio, is then used by the Transformer decoder to generate latent facial features in codebook space in an autoregressive manner. Then, the decoder pre-trained in the first stage decodes these features into facial coefficients $\{f_i\}_{i=1}^{T_v}$ of the 3DMM representation. In contrast to the training stage where randomly sampled and temporally repeated reference facial coefficients are used as input, the mapping network of the lip branch in the inference stage takes the sequence of facial coefficients generated by the face branch and maps the audio feature to lip coefficients $\{m_i\}_{i=1}^{T_v}$. This conditioning on the facial prior ensures that the generated lip-related coefficients are more compatible with the generated facial coefficients. Ultimately, the final 3DMM representation is obtained by combining the predicted lip-related coefficients $\{m_i\}_{i=1}^{T_v}$ with the facial coefficients $\{f_i\}_{i=1}^{T_v}$ (head pose, expressions), and we use off-the-shelf image renderer from [44] to produce the real video.

4 Experiments

4.1 Experimental Setup

Dataset. We train and evaluate our method on MEAD [35], a high-quality emotional audio-visual dataset with recorded videos of different actors speaking with 8 different emotions, where each emotion has 3 intensity levels except for neutral. All the videos are converted to 25 fps and the audio sample rate is set to be 16kHz.

Evaluation Metrics. To assess the texture quality of generated videos, **SSIM** calculates the structural similarity between reference and generated images, while **CPBD** computes the cumulative probability of blur detection for the sharpness of generated frames. To evaluate lip syncronization, we adopt the confidence score $\textbf{Sync}_{conf}$ of SyncNet [6] and Landmark Distance around mouth (**M-LMD**) proposed in [3]. To evaluate the accuracy of generated facial expressions, we adopt the Landmark Distance on the whole face (**F-LMD**). Following [11], we use a pre-trained emotion classifier from [32] to calculate emotion

accuracy $\textbf{Emo}_{Acc}$ of generated talking head videos. Since the facial expression should have continuous emotion intensities instead of one single hard label for the whole video, we calibrate the emotion label based on the real video's emotion per 5 frames as ground truth (i.e., Emo_{Acc} is 100% for real video) rather than one single categorical label in MEAD [35] for the entire duration.

5 Implementation Details

Quantization strategy. The VQ-VAE's naive training suffers from codebook collapse, as noted in previous work [30,33]. To improve codebook utilization, we employ two quantization strategies from [30]: exponential moving average (EMA) and codebook reset (Code Reset). The EMA strategy enables the codebook $\mathcal{C}$ to evolve smoothly over time with the formula $\mathcal{C}_t \leftarrow \lambda\mathcal{C}_{t-1} + (1-\lambda)\mathcal{C}_{t-1}$, where $\mathcal{C}_t$ represents the codebook at iteration t and λ is the exponential moving constant. On the other hand, the Code Reset strategy identifies inactive codes and reassigns them during training.
VQ-VAE. The temporal downsampling and upsampling rate l is 4 of VQ-VAE autoencoder, and the size K of learnable codebook is 1024 of dimension 256. We train the VQ-VAE with AdamW [21] optimizer with warm up learning rate, and with $[\beta_1, \beta_2] = [0.9, 0.99]$, exponential moving constant $\lambda = 0.99$ and batch size 512.
Cross-modal Transformer. The audio feature is extracted by HuBERT model [12] which finetunes on downstream emotion recognition task [41][1]. The context window size h of self-attention temporal bias mask is 4, and the temporal alignment ratio k of cross-attention alignment mask is set to 8. The second stage training is optimized with AdamW [21] with a learning rate $1e-4$ and batch size of 128.
Lip branch. The model is trained with continuous 5 frames and the audio feature for each frame is a 0.2s mel-spectrogram.

5.1 Quantitative and Qualitative Evaluation

Comparison Approaches. We compare our work with MakeItTalk [47], Wav2Lip [27], Audio2Head [36], SadTalker [44]. We chose these methods as comparison because (i) both the source code and pre-trained model weights are readily accessible; (ii) they represent the state-of-the-art talking-head video generation using arbitrary identities, which means no need for additional training or fine-tuning for a new identity that was not encountered during the training process; (iii) they take only identity reference image and audio as input for inference without any extra emotion label or driving frames.
Evaluation Results. Tab. 1 showcases our model outperforms previous approaches in terms of both landmark alignment F-LMD and M-LMD, and emotion accuracy Emo_{Acc} calibrated with real video. Additionally, our model exhibits comparable performance to other fully talking-head generation methods

[1] https://huggingface.co/superb/hubert-base-superb-er

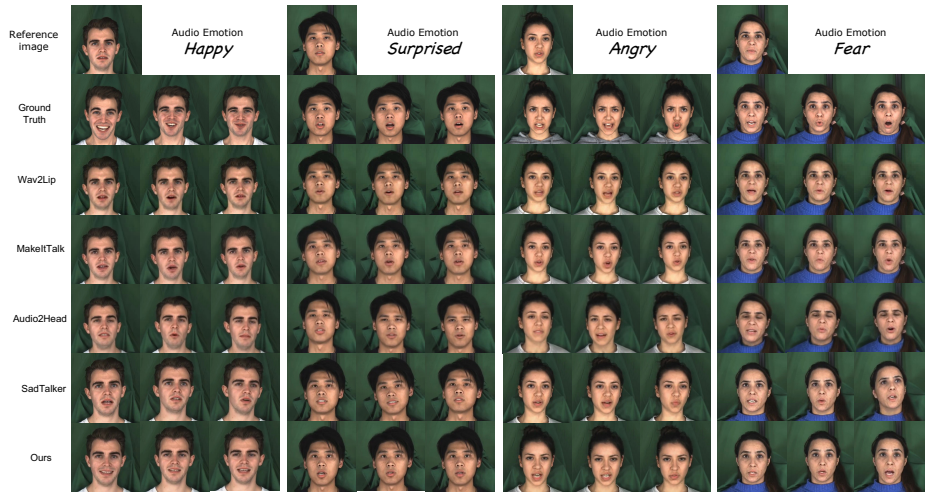

Fig. 3: Qualitative results of generated talking head frames given a neutral reference image and audio with different emotions. Note that all the models only take a reference image and audio as input, without any explicit emotion labels.

in terms of lip synchronization metrics. Note that Wav2Lip [27] focuses solely on lip generation and applies a copy-paste technique to insert the generated lip into the original image, resulting in the highest CPBD score, and it achieves the highest Sync_{conf} score since it's optimized with SyncNet [6] as a discriminative loss. In addition, this model-based metric's another limitation is that it fits too closely to the distribution of another dataset, resulting in scores that are even better than those of the real video in the MEAD [35] dataset.

The qualitative results of the generated video frames can be seen in Fig. 3. All models use a neutral reference image and emotionally intense audio as input. Notably, our model produces a diverse range of facial expressions that align with the emotional content conveyed by the audio without relying on explicit emotion labels. Please refer to the webpage for more generated videos (different emotion intensities, in-the-wild generation, and flexible control).

User Study. Automatic evaluation metrics cannot adequately measure the naturalness and subtle expressions of generated videos. Therefore, we conducted a human evaluation on Amazon MTurk by generating 100 videos and comparing our model with runner-ups. The details and interface for the evaluation can be found in the supplementary. The outcomes are illustrated in Fig. 4 (b), where our model outperforms other methods across all evaluation facets. Notably, emotion alignment with audio attained the highest count of favorable ratings. In contrast to the automatic metrics presented in Tab. 1, our model achieves superior lip synchronization performance compared to Wav2Lip [27]. This difference can be attributed to the fact that users also take factors such as blurriness and the alignment of lip movements with audio fluctuations into consideration.

Method	F-LMD ↓	M-LMD ↓	Sync_{conf} ↑
SadTalker ≡ ExpNet $\mathbb{R}^{64}$ + PoseVAE $\mathbb{R}^{6}$	4.764	5.620	2.120
Our face branch $\mathbb{R}^{57}$ + ExpNet (lip) $\mathbb{R}^{13}$	4.544	5.231	2.473
Emotalker (Our face branch $\mathbb{R}^{57}$ + Our lip branch $\mathbb{R}^{13}$)	4.348	5.139	2.304

Table 2: Ablation on how our separated face and lip branches contribute to the performance.

5.2 Ablation Studies

We further investigate the impact of our separated face and lip branches on the overall performance. In the baseline SadTalker [44] (first line) in Tab. 2, the 3DMM coefficients are separated into expression $\in \mathbb{R}^{64}$ generated by ExpNet and pose $\in \mathbb{R}^{6}$ generated by PoseVAE. The second line combines facial (non-lip) coefficients $\in \mathbb{R}^{57}$ generated by our face branch with lip-related coefficients generated by ExpNet from SadTalker [44], which shows the effectiveness of our proposed VQ-VAE and cross-modal Transformer in enhancing emotion-aware facial motion synthesis. The third line is EmoTalker default setting including our proposed separate face and lip branch synthesis, which further enhances the M-LMD score. The impact of the temporal downsampling rate and the additional classification loss ($\mathcal{L}_{cls}$) on the reconstruction quality of VQ-VAE is presented in Tab. 3. The result indicates additional $\mathcal{L}_{cls}$ enhances the performance by facilitating the differentiation of various facial motion representations. Moreover, we conducted an ablation study to examine the impact of three components in Transformer training: the self-attention temporal bias mask $\mathbf{B}^{\hat{F}}$, the cross-modal alignment mask $\mathbf{M}$, and the contrastive loss $\mathcal{L}_{contra}$, shown in Tab. 4.

Method	MSE ↓	Codebook Size	MSE ↓
Temporal ↓ $l = 4$ (default)	1.3×10^{-3}	512	1.69×10^{-3}
Temporal ↓ $l = 2$	1.6×10^{-3}	1024	1.32×10^{-3}
w/o $\mathcal{L}_{cls}$	2.1×10^{-3}	2048	1.37×10^{-3}

Table 3: Ablation on the effect of additional emotion classification loss $\mathcal{L}_{cls}$, temporal downsampling rate l to reconstruction MSE of VQ-VAE and codebook size.

5.3 Analysis

We further visualize the learned codebook's emotion distribution in Fig. 4 (a). After the first training stage, we encode 3DMM features extracted from each video clip to find the nearest code embedding. Then, we can associate specific codes with particular emotions. Since certain emotions may share the same

Method	F-LMD ↓	Emo_{Acc} ↑
EmoTalker (default)	4.348	0.425
EmoTalker w/o bias $\mathbf{B}^{\hat{F}}$	4.729	0.407
EmoTalker w/o alignment $\mathbf{M}$	4.802	0.413
EmoTalker w/o $\mathcal{L}_{contra}$	4.964	0.392

Table 4: Ablation on the effect of self attention temporal bias mask $\mathbf{B}^{\hat{F}}$, cross attention alignment mask $\mathbf{M}$ and contrastive loss $\mathcal{L}_{contra}$ in the Transformer training.

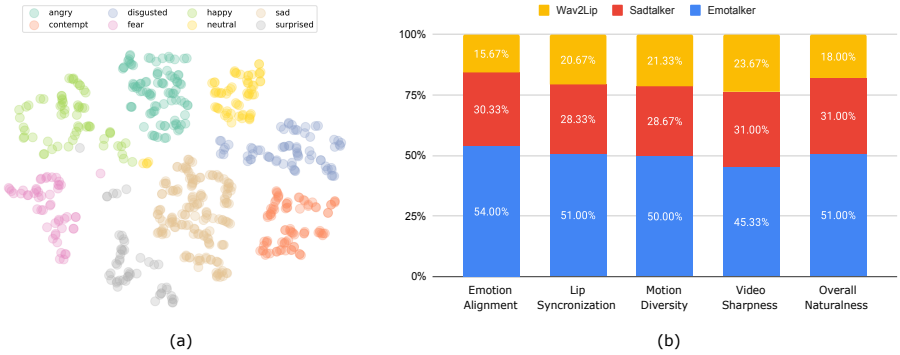

Fig. 4: a) Visualization of the emotional distribution in the learned codebook's latent representation. b) User study result.

code representation, we filter out the ambiguous indices and visualize the non-overlapping instances by utilizing t-SNE [23] to map the high-dimension feature into 2 dimension vectors. This finding indicates that the first stage training procedure enhances the model's ability to comprehend and differentiate between different emotions. Furthermore, we visualize the reconstructed centroid codes of emotionally clustered embedding in the learned codebook. As Fig. 5 shows, each centroid displays a distinct emotional expression. Please refer to the supplementary for more details.

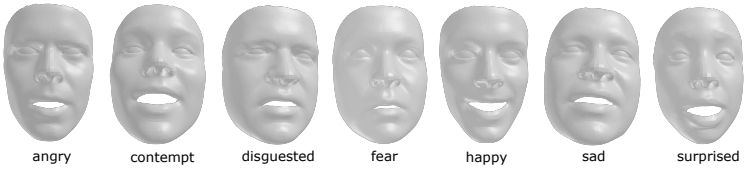

Fig. 5: Reconstructed centroid codes of emotional clustered embedding in the learned codebook.

5.4 Compare with emotional approach

All the previous emotional talking head approaches rely on driving frames to generate facial expressions that ignore the emotional signals conveyed by the audio itself. Our motivation is to model facial expressions from input audio without additional emotion information. Therefore, our setting stands apart as it exclusively utilizes audio input and doesn't require extra emotional driving frames. Therefore, it is unfair to compare with those methods.

However, we enable a fair comparison with previous emotion-based works with two scenarios. Firstly, we fit the previous method to our setting by making all the driving frames to be the neutral identity image. Second, we adopt our model for the previous setting by taking both audio and emotional driving frames as input to generate videos. The comparison result is shown in Tab. 5. We compare with EAMM [13] since EVP[2] provides only two separate pre-trained models of two target persons. As Tab. 5 shows, our method achieves comparable performance when input with driving frames with EAMM [13]. And when w/o input driving frames, our method outperforms it in each aspect, which indicates our model captures the emotional signals from audio without the help of external emotional information.

Method	w/ driving frames	Texture Quality		Landmark Alignment		Lip Sync	Emotion
		SSIM ↑	CPBD ↑	F-LMD ↓	M-LMD ↓	$Sync_{conf}$ ↑	Emo_{Acc} ↑
EAMM [13]	✓	0.665	0.316	2.541	3.672	1.837	0.718
EmoTalker (ours)	✓	0.695	0.372	2.363	3.842	2.538	0.762
EAMM [13]	×	0.660	0.307	7.985	5.133	1.605	0.183
EmoTalker (ours)	×	0.705	0.362	4.348	5.139	2.304	0.425

Table 5: Compare with emotional talking head work EAMM [13]. In the upper section, we adapt our model to the previous setting by taking both audio and emotional driving frames as input to generate videos. In the lower section, we fit EAMM to our setting by making all the driving frames to be the neutral identity image.

6 Conclusion

In conclusion, we have presented EmoTalker, a novel approach for talking head synthesis that generates emotion-aware facial expressions and lip movements from audio without relying on reference frames or emotion labels. Our method uses 3D motion coefficient representation for facial motions and models lip-related coefficients and the remaining separately. For lip movement, we start with a pre-trained audio encoder and map it to lip coefficients. For the facial expression branch, we leverage a two-stage training strategy that includes an emotion-aware facial expression prior learning and a cross-modal Transformer to explicitly model the correlation between audio and facial motions. Our approach outperforms previous methods and generates more realistic talking head videos with synchronized lip movements and natural facial expressions aligned with emotionally intensive audio.

[2] https://github.com/jixinya/EVP

References

1. Ao, T., Gao, Q., Lou, Y., Chen, B., Liu, L.: Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. ACM Transactions on Graphics (TOG) **41**(6), 1–19 (2022) 4
2. Chen, L., Cui, G., Liu, C., Li, Z., Kou, Z., Xu, Y., Xu, C.: Talking-head generation with rhythmic head motion. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX. pp. 35–51. Springer (2020) 1, 3
3. Chen, L., Li, Z., Maddox, R.K., Duan, Z., Xu, C.: Lip movements generation at a glance. In: Proceedings of the European conference on computer vision (ECCV). pp. 520–535 (2018) 9
4. Chen, L., Maddox, R.K., Duan, Z., Xu, C.: 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) 1, 3
5. Cheng, K., Cun, X., Zhang, Y., Xia, M., Yin, F., Zhu, M., Wang, X., Wang, J., Wang, N.: Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022) 1
6. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13. pp. 251–263. Springer (2017) 9, 11
7. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019) 3, 4
8. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021) 4
9. Fan, Y., Lin, Z., Saito, J., Wang, W., Komura, T.: 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) 6
10. Guo, Y., Chen, K., Liang, S., Liu, Y.J., Bao, H., Zhang, J.: Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5784–5794 (2021) 3
11. Gururani, S., Mallya, A., Wang, T.C., Valle, R., Liu, M.Y.: Spacex: Speech-driven portrait animation with controllable expression. arXiv preprint arXiv:2211.09809 (2022) 9
12. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3451–3460 (2021) 6, 7, 10
13. Ji, X., Zhou, H., Wang, K., Wu, Q., Wu, W., Xu, F., Cao, X.: Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022) 3, 4, 14
14. Ji, X., Zhou, H., Wang, K., Wu, W., Loy, C.C., Cao, X., Xu, F.: Audio-driven emotional video portraits. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14080–14089 (2021) 4

15. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 4
16. KR, P., Mukhopadhyay, R., Philip, J., Jha, A., Namboodiri, V., Jawahar, C.: Towards automatic face-to-face translation. In: Proceedings of the 27th ACM international conference on multimedia. pp. 1428–1436 (2019) 1
17. Kreuk, F., Polyak, A., Copet, J., Kharitonov, E., Nguyen, T.A., Rivière, M., Hsu, W.N., Mohamed, A., Dupoux, E., Adi, Y.: Textless speech emotion conversion using discrete and decomposed representations. arXiv preprint arXiv:2111.07402 (2021) 7
18. Lakhotia, K., Kharitonov, E., Hsu, W.N., Adi, Y., Polyak, A., Bolte, B., Nguyen, T.A., Copet, J., Baevski, A., Mohamed, A., et al.: On generative spoken language modeling from raw audio. Transactions of the Association for Computational Linguistics **9**, 1336–1354 (2021) 7
19. Liang, B., Pan, Y., Guo, Z., Zhou, H., Hong, Z., Han, X., Han, J., Liu, J., Ding, E., Wang, J.: Expressive talking head generation with granular audio-visual control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3387–3396 (2022) 3
20. Liu, X., Wu, Q., Zhou, H., Du, Y., Wu, W., Lin, D., Liu, Z.: Audio-driven co-speech gesture video generation. arXiv preprint arXiv:2212.02350 (2022) 4
21. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 10
22. Ma, Y., Wang, S., Hu, Z., Fan, C., Lv, T., Ding, Y., Deng, Z., Yu, X.: Styletalk: One-shot talking head generation with controllable speaking styles. arXiv preprint arXiv:2301.01081 (2023) 2, 3, 5
23. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008) 13
24. Ng, E., Joo, H., Hu, L., Li, H., Darrell, T., Kanazawa, A., Ginosar, S.: Learning to listen: Modeling non-deterministic dyadic facial motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20395–20405 (2022) 4
25. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: 2009 sixth IEEE international conference on advanced video and signal based surveillance. pp. 296–301. Ieee (2009) 5
26. Polyak, A., Adi, Y., Copet, J., Kharitonov, E., Lakhotia, K., Hsu, W.N., Mohamed, A., Dupoux, E.: Speech resynthesis from discrete disentangled self-supervised representations. arXiv preprint arXiv:2104.00355 (2021) 7
27. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 484–492 (2020) 1, 2, 3, 8, 9, 10, 11
28. Press, O., Smith, N.A., Lewis, M.: Train short, test long: Attention with linear biases enables input length extrapolation. arXiv preprint arXiv:2108.12409 (2021) 6
29. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021) 4
30. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems **32** (2019) 10

31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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) 4
32. Siqueira, H., Magg, S., Wermter, S.: Efficient facial feature learning with wide ensemble-based convolutional neural networks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 5800–5809 (2020) 9
33. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017) 4, 10
34. Vougioukas, K., Petridis, S., Pantic, M.: Realistic speech-driven facial animation with gans. International Journal of Computer Vision **128**, 1398–1413 (2020) 3
35. Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., Loy, C.C.: Mead: A large-scale audio-visual dataset for emotional talking-face generation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI. pp. 700–717. Springer (2020) 3, 4, 5, 9, 10, 11
36. Wang, S., Li, L., Ding, Y., Fan, C., Yu, X.: Audio2head: Audio-driven one-shot talking-head generation with natural head motion. arXiv preprint arXiv:2107.09293 (2021) 1, 3, 9, 10
37. Wiles, O., Koepke, A., Zisserman, A.: X2face: A network for controlling face generation using images, audio, and pose codes. In: Proceedings of the European conference on computer vision (ECCV). pp. 670–686 (2018) 1, 3
38. Williams, W., Ringer, S., Ash, T., MacLeod, D., Dougherty, J., Hughes, J.: Hierarchical quantized autoencoders. Advances in Neural Information Processing Systems **33**, 4524–4535 (2020) 4
39. Wu, H., Jia, J., Wang, H., Dou, Y., Duan, C., Deng, Q.: Imitating arbitrary talking style for realistic audio-driven talking face synthesis. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1478–1486 (2021) 3
40. Xing, J., Xia, M., Zhang, Y., Cun, X., Wang, J., Wong, T.T.: Codetalker: Speech-driven 3d facial animation with discrete motion prior. arXiv preprint arXiv:2301.02379 (2023) 4, 6, 8
41. Yang, S.w., Chi, P.H., Chuang, Y.S., Lai, C.I.J., Lakhotia, K., Lin, Y.Y., Liu, A.T., Shi, J., Chang, X., Lin, G.T., et al.: Superb: Speech processing universal performance benchmark. arXiv preprint arXiv:2105.01051 (2021) 10
42. Ye, Z., Jiang, Z., Ren, Y., Liu, J., He, J., Zhao, Z.: Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. arXiv preprint arXiv:2301.13430 (2023) 3
43. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations. arXiv preprint arXiv:2301.06052 (2023) 4
44. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. arXiv preprint arXiv:2211.12194 (2022) 2, 8, 9, 10, 12
45. Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X.: Talking face generation by adversarially disentangled audio-visual representation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 9299–9306 (2019) 3
46. Zhou, H., Sun, Y., Wu, W., Loy, C.C., Wang, X., Liu, Z.: Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4176–4186 (2021) 1, 3

47. Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., Li, D.: Makelttalk: speaker-aware talking-head animation. ACM Transactions On Graphics (TOG) **39**(6), 1–15 (2020) 1, 2, 3, 9, 10