MusicFace: Music-driven Expressive Singing Face Synthesis

Pengfei Liu[1] Wenjin Deng[1] Hengda Li[1] Jintai Wang[1] Yinglin Zheng[1] Yiwei Ding[1]
Xiaohu Guo[2] Ming Zeng[1]
[1]Xiamen University [2]The University of Texas at Dallas

Abstract

It is still an interesting and challenging problem to synthesize a vivid and realistic singing face driven by music signal. In this paper, we present a method for this task with natural motions of the lip, facial expression, head pose, and eye states. Due to the coupling of the mixed information of human voice and background music in common signals of music audio, we design a decouple-and-fuse strategy to tackle the challenge. We first decompose the input music audio into human voice stream and background music stream. Due to the implicit and complicated correlation between the two-stream input signals and the dynamics of the facial expressions, head motions and eye states, we model their relationship with an attention scheme, where the effects of the two streams are fused seamlessly. Furthermore, to improve the expressiveness of the generated results, we propose to decompose head movements generation into speed generation and direction generation, and decompose eye states generation into the short-time eye blinking generation and the long-time eye closing generation to model them separately. We also build a novel SingingFace Dataset to support the training and evaluation of this task, and to facilitate future works on this topic. Extensive experiments and user study show that our proposed method is capable of synthesizing vivid singing face, which is better than state-of-the-art methods qualitatively and quantitatively.

Keywords: Music-driven face generation, singing face, generative adversarial network.

1. Introduction

With the advancement of computer vision and computer graphics, synthesizing vivid and realistic dynamic face is becoming possible and has been attracting more and more attention from CV/CG communities. Recent progresses [13, 43, 7, 57, 61, 23, 45, 57] show the great potential of this topic in a variety of applications, such as human-computer interaction [34, 58], video making [37, 59, 63, 40], and news anchor composition[51, 47], *etc.*

Despite the recent progresses of dynamic face synthesis [13, 43, 7, 57, 61, 23, 45, 66, 24] and its potential applications [34, 58, 37, 59, 63, 40, 51, 47], it is still an open problem regarding how to synthesize a vivid face as expressive as possible.

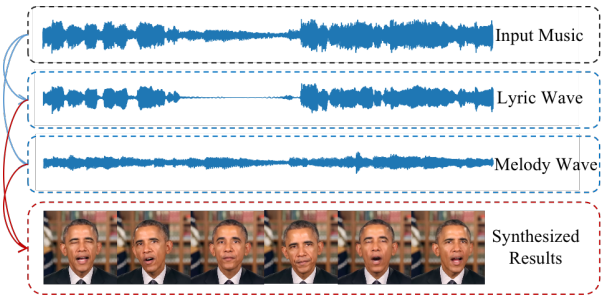


Figure 1. **Problem description.** Our goal is to synthesize a vivid dynamic singing face coherent with the input music audio, which is mixed with human voice and background music.

Existing work in the literature focuses on generating coherent dynamics of faces according to input speech audio [4, 38, 13, 43, 36, 46, 52, 7, 57, 61]. However, in many emotional scenarios, it is required that the head synthesis is driven by a composite audio which is coupled with not only speech but also other signals, *e.g.*, the music audio contains both human voice and background music signals. Therefore, in this paper, we investigate the problem of synthesizing a vivid dynamic face which is not only in-sync but also delivers coherent facial dynamics with the input music audio, as is illustrated in Fig. 1. This is a non-trivial task, which can not be handled directly by existing methods. This is because common music audios are mixed by coupled human voice and background music signals, while most of the existing methods are designed for synthesizing face according to only the human speech signals, which will lead to undesired results due to the entanglement of different audio signals.

To tackle this challenge, we investigate the implicit correlation between the input signals and the facial dynamics. We treat the input music audio as a mixed signal which includes a human voice signal and a background music signal. According to previous work [43, 36, 46, 52, 7, 57, 61] and our observation, we argue that the lip movement is majorly related to the voice signal (also called the speech channel),

while the head pose, facial expression, eye states relate to both the voice signal and background music signal. However, we would like to ask the questions: *Are these subjective observations true?* and *How much do the human voice and background music signals affect the face dynamics?* To answer these questions, we devise a decouple-and-fusion framework for this task. Firstly, we separate the input music audio into the human voice channel and the background music channel. Then we dynamically fuse these two separated signals in a feature selection fashion by introducing a *Attention-based Modulator*. The *Attention-based Modulator* modulates and balances the two signals for the downstream generators of facial expressions, head motions, and eye states.

In the singing scenarios, the motions of the head and eyes are usually emotional and dramatic, which raises challenges for generators to learn the more diverse and expressive motions as compared with the previous talking scenarios. We propose two ingredients to improve the expressiveness of the synthesis result. For the movement of the head, we propose to learn the rhythm of head motion that is decoupled from the absolute moving velocity, thus factoring off the ambiguity of the mapping between audio and head movement. For the eye states, we propose to synthesize both eye blinking and long-time eye closing states, which delivers much more expressiveness as compared with previous methods.

Besides, to learn the complex and implicit relationship between the music audio and face dynamics, we build a SingingFace Dataset from our recordings. The dataset contains over 600 singing videos with synchronous music audio. To our best knowledge, this is the first dataset regarding face dynamics and music audio. We believe it will promote future research on this topic.

In summary, this paper is featured as follows:

- This is the first framework for synthesizing a singing face video driven by the input music audio mixed with human voice and background music signals. In the framework, we introduce the *Attention-based Modulator* to balance the effects of the two signals on the head movements, expressions, and eye states.
- We propose to synthesize the speed and direction of head movements separately, instead of predicting head pose directly. The simple-yet-effective modification leads to more consistent head dynamics in line with music rhythm. Besides, we propose to decompose the eye states into eye blinking and long-time eye closing, which is much more realistic in singing scenarios.
- We build the first dataset which contains expressive singing face videos with synchronous music audio, and make it public to facilitate future research on this topic.

[*]https://vcg.xmu.edu.cn/datasets/singingface/index.html

2. Related Work

2.1. Audio-driven Talking Face Synthesis

Audio-driven face synthesis has been widely explored. Previous work [3, 16, 48, 38, 13] focuses on establishing the mapping between facial motion factors and audio features. Brand [3] uses a Hidden Markov Model (HMM) to predict facial motions. Ezzat *et al.* [16] leverage an example-based method mapping phonemes to mouth shape and texture parameters in the Principle Component Analysis (PCA) space. Wang *et al.* [48] attempt to model a mapping between Mel-Frequency Cepstral Coefficients (MFCC) and PCA model parameters via an HMM approach. Benefiting from deep learning techniques, some works have been proposed to generate more diverse faces in sync with input audio. Shimba *et al.* [38] estimate active appearance model (AAM) parameters with the Long Short-Term Memory (LSTM) network. Cudeiro *et al.* [13] employ convolutions to encode speech and decode facial attributes to animate a 3D template.

Several methods [4, 36, 46, 9, 43, 14, 17, 65, 55, 18, 53] merely synthesize facial region texture with lip-synced motions. Among the above approaches, a broad class of them generate identity-preserving face with static head pose using GANs [9, 14, 65]. Other methods synthesize lip-synced texture of mouth, then rewrite the mouth area of source frames according to the input audio [43, 46, 36, 4, 53] or text [17, 55]. However, due to the dependence on the original video, they can only generate limited head poses. To address this problem, Chen *et al.* [7], Yi *et al.* [57] and Zhang *et al.* [60] estimate head movements from input audio. Most recently, Zhang *et al.* [62], Li *et al.* [30] and Guo *et al.* [18] synthesize photo-realistic 3D head with natural head poses and synchronized lip motions using popular neural rendering techniques. Wang *et al.* [49, 50] even generate photo-realistic faces from one-shot reference image with natural motions.

2.2. Music-driven Animation

Music-driven human pose animation has been studied for decades. Early work [6, 28, 39] formulate the task as a template matching problem. Lee *et al.* [28] and Shiratori *et al.* [39] generate dance motion sequences with musical similarity based on manually defined audio features, while Cardle *et al.* [6] edit motions guided by musical features. Due to the limitations of capacity, these template matching approaches are not competent to generate diverse and natural dance motions.

With the great success of deep neural network, more researchers address the music-to-dance as a generation problem with learning-based techniques. Recent methods employ auto encoder-decoder [27], LSTM [1, 44, 54, 68, 25], GAN [26, 42], and Transformer [21, 29, 31]. Even though

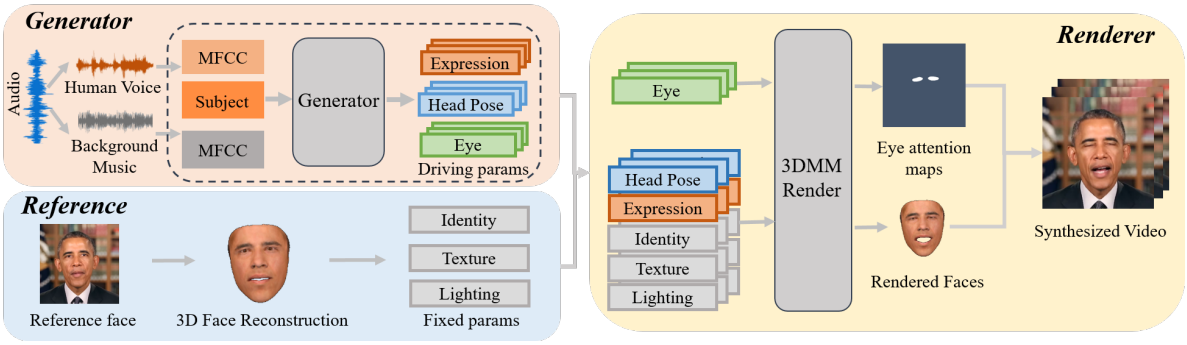


Figure 2. **Framework overview.** Taking human voice and background music separated from music audio as input, the Generator module generates facial driving parameters (expressions, head poses and eye states). Conditioned with fixed parameters (identity, texture, lighting) extracted from a reference face image and the driving parameters, the Renderer module aims to synthesize a photo-realistic video. Specifically, eye state parameters are encoded into eye attention maps, and other parameters provide a 3D model guidance to render faces. Finally, an expressive and rhythmic singing face video is rendered by combining rendered faces with eye attention maps.

some work [26, 56] apply action units to further explore the correlations between pose and music, it is still challenging to generate diverse, rhythmic and expressive dance motion.

It is interesting to note that music-driven singing face synthesis remains a rarely studied open problem. Song2Face [22] is the only one designed for singing scenarios up to now to the best of our knowledge. However, it operates on plain human singing voice, only working well without the disturbance of background music. The decouple-and-fuse framework presented in this paper can generate realistic and rhythmical facial dynamics from mixed music wave. Therefore, the paper will open novel research directions in the domain of music-guided person synthesis.

3. Methodology

3.1. Problem Definition

In previous researches [9, 66, 57, 62, 41, 67], given a piece of speech audio $\mathcal{A}$ and a short reference video (or a single face image) $\mathcal{V}$, the ultimate goal is to generate a realistic talking face video $\mathcal{S}$ synchronized with the input audio $\mathcal{A}$, which can be represented as:

$$\begin{aligned} \mathcal{F}_{exp}, \mathcal{F}_{pose}, \mathcal{F}_{eye} &= \mathbf{G}(\mathbf{E}(\mathcal{A})), \\ \mathcal{S} &= \mathbf{R}(\mathcal{F}_{exp}, \mathcal{F}_{pose}, \mathcal{F}_{eye}, \mathcal{V}), \end{aligned} \tag{1}$$

where $\mathcal{F}_{exp}$, $\mathcal{F}_{pose}$, $\mathcal{F}_{eye}$ denote the facial expression, head pose, and eye state parameters synthesized by a generator $\mathbf{G}$, respectively. $\mathbf{E}$ refers to an audio feature extractor and $\mathbf{R}$ denotes a rendering network synthesizing photo-realistic images.

However, directly predicting driving parameters from audio is not up to music scenarios due to the complicated mutual influences between human voice containing lyric information and background music containing melody information. We propose a decouple-and-fuse strategy to tackle the above problem, which firstly adopts an audio source separation model $\mathbf{O}$ to decompose music into human voice $\mathcal{A}^v$ and background music $\mathcal{A}^b$, then gets encoded lyric feature $\mathcal{L}$ and melody feature $\mathcal{M}$ respectively using an attention-assisted two-stream encoder $\mathbf{E}$. It encodes lyric and melody separately, and modifies the relative contribution of the two encoded features on the generation process through an attention mechanism. Finally a generator $\mathbf{G}$ is employed to generate the driving parameters of a singing face video $\mathcal{S}$ from the decoupled lyric feature and melody feature. The full pipeline can be formulated as follows:

$$\begin{aligned} \mathcal{A}^v, \mathcal{A}^b &= \mathbf{O}(\mathcal{A}), \\ \mathcal{L}, \mathcal{M} &= \mathbf{E}(\mathcal{A}^v, \mathcal{A}^b), \\ \mathcal{F}_{exp}, \mathcal{F}_{pose}, \mathcal{F}_{eye} &= \mathbf{G}(\mathcal{L}, \mathcal{M}), \\ \mathcal{S} &= \mathbf{R}(\mathcal{F}_{exp}, \mathcal{F}_{pose}, \mathcal{F}_{eye}, \mathcal{V}). \end{aligned} \tag{2}$$

As illustrated in Fig. 2, our overall framework contains three components: 1) a driving parameter **generator** to translate music audio to facial expression, head pose, and eye states, 2) a **reference** module extracting fixed parameters such as face identity given a human face, 3) and a **renderer** to synthesize photo-realistic frames conditioned on above parameters. We employ a conditional-GAN-based method as our renderer, which is of the same architecture as [62]. To enhance the expressiveness of singing faces, the generator $\mathbf{G}$ is designed as the following encoder-decoder architecture as is shown in Fig. 3. The Encoder (Sec. 3.2) consists of a Two-stream Audio Encoder (TSAE) to encode lyric and melody separately and an Attention-based Modulator (ATM) to balance the contribution of different audio features. The Decoder (Sec. 3.3) contains three down-

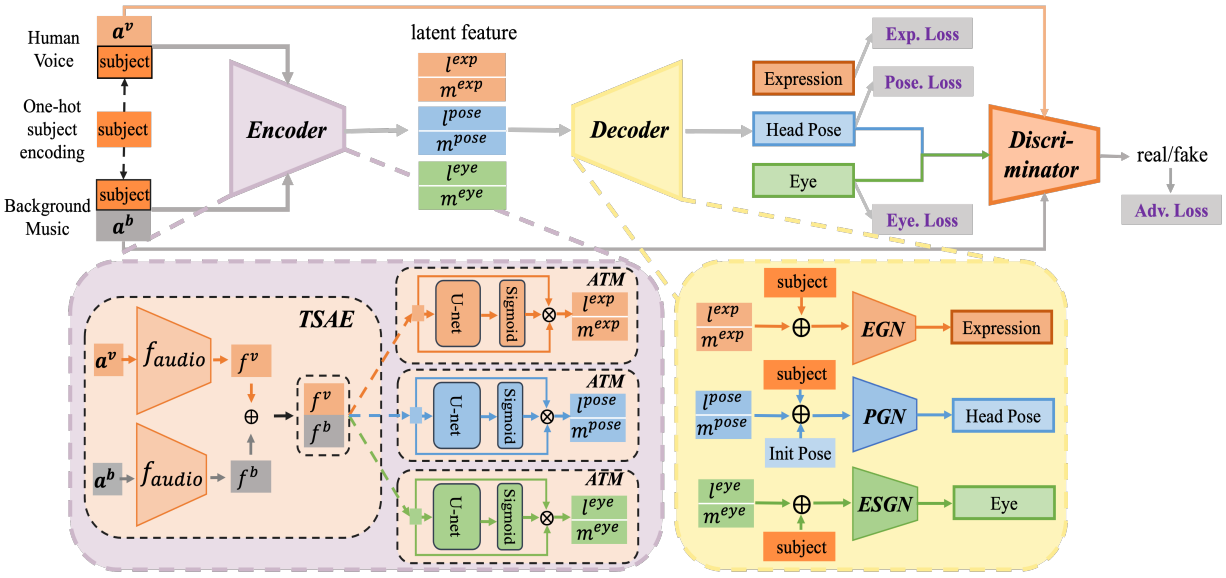


Figure 3. **The Architecture of Our Generator.** Our generator contains an Encoder and a Decoder. The Encoder consists of a Two-stream Audio Encoder (TSAE) and an Attention-based Modulator (ATM). The Decoder contains three downstream generators, including Expression Generation Network (EGN), Pose Generation Network (PGN), and Eye State Generation Network (ESGN).

stream generators, including Expression Generation Network (EGN) for the generation of facial expression parameters, Pose Generation Network (PGN) for the generation of head pose dynamics, and Eye State Generation Network (ESGN) for the generation of eye state parameters. In the next subsections, we will introduce the five essential parts respectively and provide the corresponding learning objective and training strategy.

3.2. Encoder

As mentioned above, lyric and melody entangled in the original music wave show a complicated relationship in guiding the generation process of human face dynamics, making it difficult for the generation network to synthesize vivid face dynamics directly from plain music features. To tackle the problem, we employ a decouple-and-fuse strategy. Specifically, using a state-of-the-art audio source separation model Spleeter [19], we decompose the original music into human voice and background music. Then we encode lyric from human voice and melody from background music separately using a two-steam audio encoder. Finally, we adjust them with attention-based modulators to distribute the relative contribution of lyric and melody for each specific generation task.

3.2.1 Audio Feature Extraction

Taking the separated audio wave (human voice or background music) sampled at 16KHz of T seconds as input, we extract mel-frequency cepstral coefficients (MFCC) and their first derivatives with 25ms window size and 10ms window step, resulting in 26-D audio features of 100 frames per second. Furthermore, in order to incorporate temporal information and match the frequency of video frames (30 fps), the feature sequence are converted to overlapping windows of size 39 (corresponding to 390ms) at 30 fps. Therefore, the output feature is a three-dimensional array with the size $(30 \times T, 39, 26)$.

3.2.2 Two-stream Audio Encoder (TSAE)

Given the separated human voice feature $\mathcal{A}^v$ and background music feature $\mathcal{A}^b$, we adopt a Two-stream Audio Encoder (TSAE) that consists of two networks AE^v and AE^b to encode the MFCC features of human voice a_t^v and background music a_t^b, separately:

$$\begin{aligned} f_t^v &= \mathrm{AE}^v(a_t^v), \\ f_t^b &= \mathrm{AE}^b(a_t^b), \end{aligned} \tag{3}$$

where AE^v and AE^b are 1D temporal convolutional neural networks with residual blocks sharing the same network structure , and f_t^v, f_t^b indicate the encoded audio features. The subscript t indicates time step, and the superscripts v and b indicate human voice and background music, respectively. The encoded audio features of the full audio sequence $\mathbf{f}^v$ and $\mathbf{f}^b$ are obtained after stacking the audio features of each time step.

3.2.3 Attention-based Modulator (ATM)

For a specific downstream generation task, the relative contributions of features representing different specific semantic information change over time and are even interconnected with each other. For example, image the head pose dynamics of a person singing a line of a song. He will prepare to vocalize, then sing, and shut his mouth finally. In the first and third stages, he rotates his head rhythmically dominated by melody. But when he vocalizes, melody in background music and lyric in human voice influence his head movements together. So the dominant source changes over time and even becomes ambiguous during vocalization, making the generation task difficult.

Therefore, in order to generate vivid human face movements, we introduce a channel attention mechanism similar to the attention mechanism proposed in [20] to determine the relative contribution between lyric and melody on the generation result. The only difference is that, to consider the long-time dependence between the audio features of different time steps, we select a temporal U-net to generate attention weights instead of using a simple multi layer perceptron (MLP) network. Specifically, given the separately encoded audio features, we employ an Attention-based Modulator(ATM) for each generation task to estimate an attention weight of each feature map in embedding feature $\mathbf{f^v}$ and $\mathbf{f^b}$ to adjust the relative importance between them:

$$\begin{aligned} \mathbf{att} &= \sigma(\textbf{U-net}(\mathbf{f^v} \oplus \mathbf{f^b})), \\ [\mathbf{l}, \mathbf{m}] &= \mathbf{ATM}(\mathbf{f^v} \oplus \mathbf{f^b}) \\ &= \mathbf{att} \odot (\mathbf{f^v} \oplus \mathbf{f^b}), \end{aligned} \tag{4}$$

where $\mathbf{l}$ and $\mathbf{m}$ denote the final output embedding of lyric and melody features for the full audio sequence respectively, $\oplus$ represents the concatenate operation on the feature channel dimension, and $\odot$ indicates the element-wise product. **ATM** indicates the Attention-based Modulator implemented using an temporal u-net network **U-net** and σ represents the sigmoid activation function.

As shown in Fig. 3, we employ one **ATM** to learn the optimal attention weight for each downstream task. Specifically, we apply a total of three **ATM**s on $\mathbf{f^v}$ and $\mathbf{v^b}$, to get $\mathbf{l^{exp}}$ and $\mathbf{m^{exp}}$ for expression generation task, $\mathbf{l^{pose}}$ and $\mathbf{m^{pose}}$ for head pose generation task, and $\mathbf{l^{eye}}$ and $\mathbf{m^{eye}}$ for eye state generation task, respectively.

3.2.4 Subject Style Embedding

Our TSAE, EGN, PGN, and ESGN are conditioned on the subject code to learn subject-specific styles, adopting a similar strategy in [13], which encodes each subject in the dataset using a one-hot subject encoding. At training stage, the subject encoding is concatenated to each input MFCC feature a_t^v and a_t^b, and also concatenated to the final output l_t and m_t of the ATM.

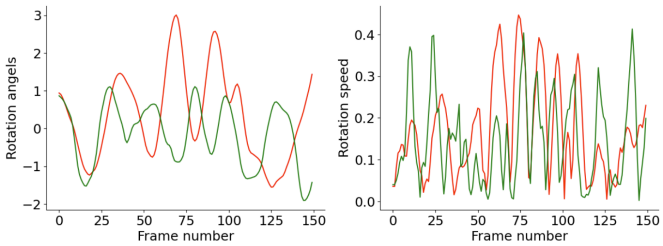


Figure 4. **The Euler angle (R_y) dynamics of a person singing the same song twice.** It shows although the head may rotate in opposite direction, the speed still keeps similar. This observation is also valid for other head pose parameters.

3.3. Decoder

3.3.1 Expression Generation Network

We employ a simple MLP consisting of two fully connected layers and one ReLU activation layer to regress facial expression (including lip motion) parameters from the encoded lyric and melody features. The process can be formulated as:

$$\hat{f}_t = \varphi_{exp}(l_t^{exp} \oplus m_t^{exp}), \tag{5}$$

where $\hat{f}_t$ denotes the predicted facial expression parameter at time step t and φ_{exp} means the MLP for expression generation.

3.3.2 Pose Generation Network

Traditional audio-driven pose generation methods directly regress head pose parameter sequences from audio features [7, 57, 62], which does not agree with the fact that given a fixed audio sequence, different people even the same person singing the same song multiple times can produce mostly different head pose sequences as shown in Fig. 4.

We find that although the dynamics of head pose vary when the same person sings the same song multiple times, as shown in Fig. 4, the speed of head pose keeps similar in line with the rhythm of the music. Motivated by this, we propose to generate the moving speed and moving direction of the head separately, and combine them to generate head pose $p \in \mathbb{R}^6$ including Euler angles and a 3D translation vector at each time step.

Moving Speed Generation: At the first stage, we use an MLP network φ_{speed} to predict the speed of head pose parameters according to the encoded audio features at current time t:

$$\hat{s}_t = ABS(\varphi_{speed}(l_t^{pose} \oplus m_t^{pose})), \tag{6}$$

where l^{pose} and m^{pose} are the lyric and melody embedding features for pose generation, $\hat{s}_t \in \mathbb{R}^6$ is the output head

speed at time step t. As $\hat{s}_t$ can not be negative, we apply absolute function ABS to the output of φ_{speed}.

Moving Direction Generation: We use an LSTM network followed by a fully connected layer φ_{direc} to generate the direction of head movements from the encoded audio features concatenated with the previous head pose and moving velocity at the last time step:

$$\begin{aligned} \hat{v}_{t-1} &= \hat{p}_{t-1} - \hat{p}_{t-2}, \\ o_t, c_t &= \text{LSTM}(l_t^{pose} \oplus m_t^{pose} \oplus \hat{p}_{t-1} \oplus \hat{v}_{t-1}, c_{t-1}), \\ \hat{d}_t &= tanh(\varphi_{direc}(o_t)), \end{aligned} \tag{7}$$

where $\hat{p}_{t-1}$, $\hat{p}_{t-2} \in \mathbb{R}^6$ indicate the generated head pose parameters represented by Euler angles and 3D translation parameters, $\hat{v}_{t-1} \in \mathbb{R}^6$ is the predicted head pose velocity, c_{t-1} and c_t are the cell states, o_t means the output of LSTM network, $\hat{d}_t \in \mathbb{R}^6$ is the predicted moving direction, respectively at the corresponding time step.

Head Pose Generation: Finally, the pose p_t at time step t can be directly calculated by: $\hat{p}_t = \hat{p}_{t-1} + \hat{s}_t \times \hat{d}_t$.

3.3.3 Eye State Generation Network

Traditional methods usually generate only random eye blinks from audio features [62] or noise inputs [41], ignoring some long-time eye closing phenomena in singing scenarios, *e.g.*, people may close their eyes for a long time while singing the climax of the song. We decompose the generation process of eye states into random eye blinking generation and long-time eye closing state generation. Human blinks occur randomly and can be sampled from experimental predefined random distributions, but for long-time eye closing state generation, it should be learned from data.

Random Eye Blinking Generation: The normal blinks of human show regularity regarding the average human eye blinking rate and the average inter-blink duration [62]. Accordingly, we uniformly sample the blink interval $B_i \sim \mathcal{U}(a_i, b_i)$ and blink duration $B_d \sim \mathcal{U}(a_d, b_d)$ with the empirical parameters $a_i = 1.2s, b_i = 2.0s, a_d = 0.10s, b_d = 0.45s$.

Then we generate the eye state of blink dynamics $\hat{e}^{blink} \in \{0, 1\}$ according to B_i and B_d.

Long-time Eye Closing State Generation:

We employ an MLP network φ^{eye} to generate the eye state $\hat{e}_t^{long}$ at time step t:

$$\hat{e}_t^{long} = \varphi^{eye}(l_t^{eye} \oplus m_t^{eye}). \tag{8}$$

We combine the $\hat{e}_t^{blink}$ and $\hat{e}_t^{long}$ to get the composite dynamics of eye states $\hat{e}_t$:

$$\hat{e}_t = \begin{cases} \hat{e}_t^{long}, & \text{if } \hat{e}_t^{long} > 0, \\ \hat{e}_t^{blink}, & \text{otherwise .} \end{cases} \tag{9}$$

Finally, we apply a temporal gaussian filter on $\hat{e}_t$ to get more smooth eye state dynamics.

3.4. Learning Objective

We supervise our generator with the following loss functions:

$$L_{Reg} = L_{exp} + L_{pose} + L_{eye} + L_{att}, \tag{10}$$

where L_{exp}, L_{pose} and L_{eye} are the losses for facial expression, head pose, and eye states, respectively. L_{att} is the loss term for pushing ATM to select useful feature channels. Each loss term is formulated as:

$$\begin{aligned} L_{exp} &= w_1 L_{MSE}(\mathbf{f}, \hat{\mathbf{f}}) + w_2 L_{VEL}(\mathbf{f}, \hat{\mathbf{f}}), \\ L_{pose} &= w_3 L_{MMD}(\mathbf{p}, \hat{\mathbf{p}}) \\ &\quad + w_4 L_{L_1}(ABS(\mathbf{v}), ABS(\hat{\mathbf{v}})), \\ L_{eye} &= w_5 L_{L_1}(\mathbf{e}^{\mathbf{long}}, \hat{\mathbf{e}}^{\mathbf{long}}) \\ &\quad + w_6 L_{MMD}(\mathbf{e}^{\mathbf{long}}, \hat{\mathbf{e}}^{\mathbf{long}}), \\ L_{att} &= ||\mathbf{att}^{\mathbf{exp}}||_1 + ||\mathbf{att}^{\mathbf{pose}}||_1 + ||\mathbf{att}^{\mathbf{eye}}||_1, \end{aligned} \tag{11}$$

where w_1, w_2, w_3, w_4, w_5, w_6 are balancing weights. $\mathbf{f}$, $\mathbf{p}$, $\mathbf{v}$, $\mathbf{e}^{\mathbf{long}}$ are vectors containing the time serial ground truth parameters of facial expression, head pose, head moving velocity and long-time closing eye state parameters (note that we only learn long-time closing eye dynamics from data), ranging from $t = 1, 2, ..., T$. $\hat{\mathbf{f}}$, $\hat{\mathbf{p}}$, $\hat{\mathbf{v}}$, $\hat{\mathbf{e}}^{\mathbf{long}}$ are the corresponding predicted vectors. $\mathbf{att}^{\mathbf{exp}}$, $\mathbf{att}^{\mathbf{pose}}$, $\mathbf{att}^{\mathbf{eye}}$ are the predicted attention matrices for tasks of facial expression generation, head pose generation, and eye state generation, respectively. $ABS(\mathbf{x})$ denotes taking absolute values for each elements. We only supervise the absolute speed of generated head pose dynamics here, guiding the network to generate more rhythmical head pose dynamics aligned with music. $L_{MSE}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{T}\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2$ is an L^2 norm loss term, $L_{L1}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{T}\|\mathbf{x} - \hat{\mathbf{x}}\|_1^1$ is an L^1 norm loss. $L_{VEL}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{T-1}\sum_{t=1}^{T-1} \|(x_t - x_{t-1}) - (\hat{x}_t - \hat{x}_{t-1})\|_2^2$ is the velocity loss term, and L_{MMD}[32] is the maximum mean discrepancy loss to match all orders of statistics between the prediction and ground-truth. Here we use $\mathbf{x}$ to represent the ground-truth, while using $\hat{\mathbf{x}}$ for the predicted values. In our experiments, we empirically set $w_1 = 5$, $w_2 = 50$, $w_4 = 10$, $w_5 = 5$, and set other weights to 1.0.

Furthermore, in order to improve the diversity of generation results, we use an adversarial loss to fool the discriminator D, which is defined as :

$$\begin{aligned} L_{Adv} = arg \min_G \max_D \mathbb{E}_{\mathbf{p}, \mathbf{e}^{\mathbf{long}}, \mathbf{a}}[logD(\mathbf{p}, \mathbf{e}^{\mathbf{long}}, \mathbf{a})] \\ + \mathbb{E}_{\mathbf{a}, p_0}[log(1 - D(G(\mathbf{a}, p_0), \mathbf{a}))]. \end{aligned} \tag{12}$$

The total loss function in training phase is:

$$L = \lambda_1 L_{Reg} + \lambda_2 L_{Adv}. \tag{13}$$

4. Experiments

4.1. Implementation Details

Our method is implemented with PyTorch, and all the experiments are conducted on two NVIDIA RTX 3090 GPUs. For network training, we randomly sample the frame sequence with a sliding window of 128 frames. We adopt Adam optimizer during training, with a learning rate of 0.0001 for 50 epochs. Linear learning rate decay is adopted for the last 60% epochs. The hyperparameters in Eq. (13) are $\lambda_1 = 1$ and $\lambda_2 = 0.1$, respectively.

To get vivid and photo-realistic visualization results, we train a rendering-to-video network by following FACIAL [62].

4.2. Dataset Organization

As mentioned above, popular conventional datasets only contain talking face videos that lack expressiveness. To overcome this, we build a new dataset called SingingFace. SingingFace includes more than 600 singing videos with 6 human subjects. Our supplementary video shows the learned style of different subjects when training across all the 6 human subjects.

Video Collection: We organize our dataset by recording singing videos ourselves. Specifically, we collect the singing audio set first, then the face region of the person singing the song with music played simultaneously is recorded. Finally, we automatically align each video to the corresponding music audio using SyncNet [12] to ensure audio-visual synchronization.

Audio Separation: We use a state-of-the-art audio source separation model Spleeter [19] to extract the human voice as lyric information and the background music as melody information, respectively.

3D Face Reconstruction: To automatically extract face expression parameters and head poses from a singing video, we adopt Deep3DFace [15] to extract face parameters $[\alpha, \beta, \delta, \gamma, p]$, where $\alpha \in \mathbb{R}^{80}$, $\beta \in \mathbb{R}^{64}$, $\delta \in \mathbb{R}^{80}$ are the corresponding coefficient vectors for geometry, expression and texture. $\gamma \in \mathbb{R}^{27}$ is the spherical harmonics (SH) illumination coefficients. The 3D face pose $p = [R; t]$ is represented by rotation $R \in SO(3)$ and translation $t \in \mathbb{R}^3$. The PCA basis of geometry, texture, and expression are adopted from the Basel Face Model [35] and FaceWareHouse [5].

Eye State Extraction:

We employ a state-of-the-art facial analysis system OpenFace [2] to extract action unit AU45r as the eye blink parameters. Note that we observed that the distribution of the extracted AU45r values for different people varies much, so we apply min-max normalization on AU45r for each video individually. Then we apply a time length threshold $\tau = 0.5s$ to detect the short-time blinking and long-time eye closing states separately.

Table 1. **Ablation Study.**

Method	Lip Sync		Pose Realism		Eye Realism
	AVC↑	LMD↓	CCA↑	Rough↓	CCA↑
Single-Stream	1.17	1.31	0.22	0.24	0.06
Two-Stream	1.73	1.17	0.25	0.18	0.07
With-ATM	1.87	1.10	0.29	0.07	0.11
Ours	**1.90**	**1.08**	**0.33**	**0.06**	**0.12**

Data Statistics:

We collect over 600 Chinese and English singing videos totaling 40 hours with 30 FPS. Each video contains one person singing a whole song and the average length of videos is about 4 minutes. Each video has a stable camera location and appropriate lighting conditions. We randomly split out 90% of the videos for training and 10% for testing.

4.3. Ablation Study

To verify the effectiveness of the key ingredients in our proposed method, *i.e.*, 1) the audio separation step and two-stream audio encoder (TSAE), 2) the attention modulator (ATM), and 3) the head pose generation network (PGN), we study the following variants of our method:

- **Single-Stream:** no audio source separation; a single stream audio encoder is employed to encode the MFCC feature of the mixed audio wave; no ATM; and replace our PGN with an MLP network.
- **Two-Stream:** equipped with audio source separation and TSAE; no ATM; and replace our PGN with an MLP network.
- **With-ATM:** equipped with audio source separation, TSAE and ATM, and replace our PGN with an MLP network.
- **Ours:** equipped with audio source separation, TSAE, ATM, and our proposed PGN.

We compare the above variants using the splitted test set of SingingFace. We evaluate the Audio-Visual Confidence (AVC) scores proposed in [12], and Landmark Distance (LMD) introduced in [8] for lip synchronization comparison. However, to the best of our knowledge, there are no clear metrics for evaluating the realism of generated head pose and eye closing dynamics for now, which is a subjective task and is an open question to the public. Following Zhang *et al*. [64], we employ Canonical Correlation Analysis (CCA) on the generated head pose parameters sequences and eye state sequences with the ground truth and compute the Canonical Correlation as the evaluation metric for perceptual realism. Note that to emphasize the evaluation for the rhythm of the head pose dynamics that should be in line with music, we apply Canonical Correlation Analysis on the moving speed of generated head pose sequences instead

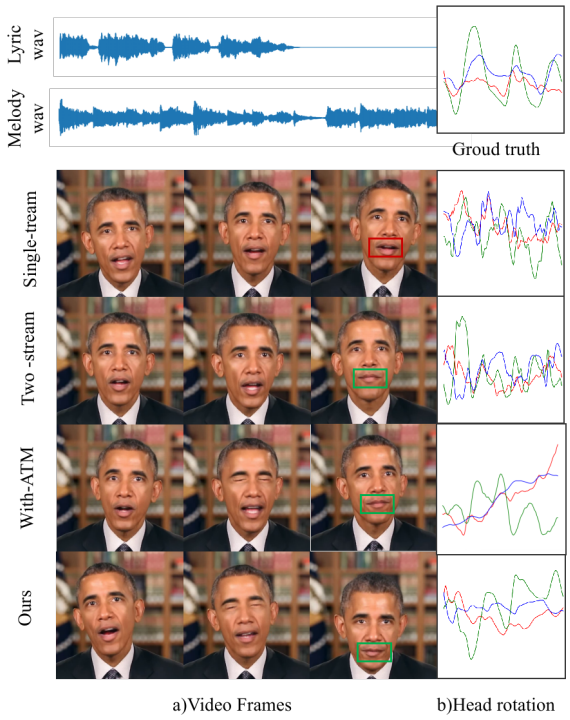


Figure 5. **Qualitative Result of Ablation Study.** It shows a) the generated frames and b) pitch (red), yaw (blue), and roll (green) of head pose dynamics. In a), the mouth generated by Single-stream still keeps open during silence (red box) while others keep closed (green box). In b), our generated head pose dynamics are smoother than others. And the turn of dominant varying angle (shown as green curve) occurs nearly at the same time with ground truth, meaning that our generated head dynamics have more similar rhythm to the ground truth recorded by a performer.

of head pose parameters themselves. We also compute the second derivative based roughness (Rough) of the generated Euler angles defined in Eq. (14) for head motion smooth evaluation:

$$Rough(R) = \frac{1}{T}\sum_{t=1}^{T} R''(t)^2, \tag{14}$$

where $R^{''}(t)$ denotes the second derivatives of head rotation angles at time step t. The quantitative results of ablation study are summarized in Tab. 1.

Effectiveness of Two Stream Design: As mentioned above, lyric and melody information are entangled together in plain music waves, making it difficult to learn facial dynamics in line with the music. It's verified that, by separating human voice and background music from plain music waves and encoding the features separately, our two-stream design greatly reduces the complexity of the lip synchronization task, thus leading to a better synchronization result. As shown in Fig. 5, if we just learn singing facial dynamics from plain music (Single-stream), the generated mouth movements are severely disrupted by the background music (*e.g.*, the mouth still keeps open during silence). On the contrary, the other variants that apply our two-stream design perform correctly. On the other hand, after applying source separation and our TSAE(Two-stream), all of the evaluation metrics have been improved a lot compared with Single-stream shown in Tab.1. This improvement can be more clearly seen in our supplementary video.

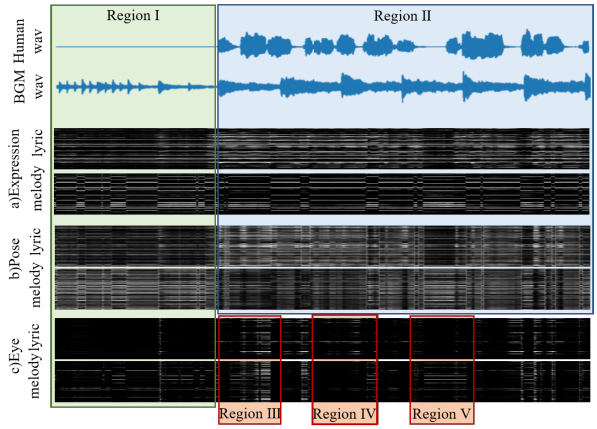
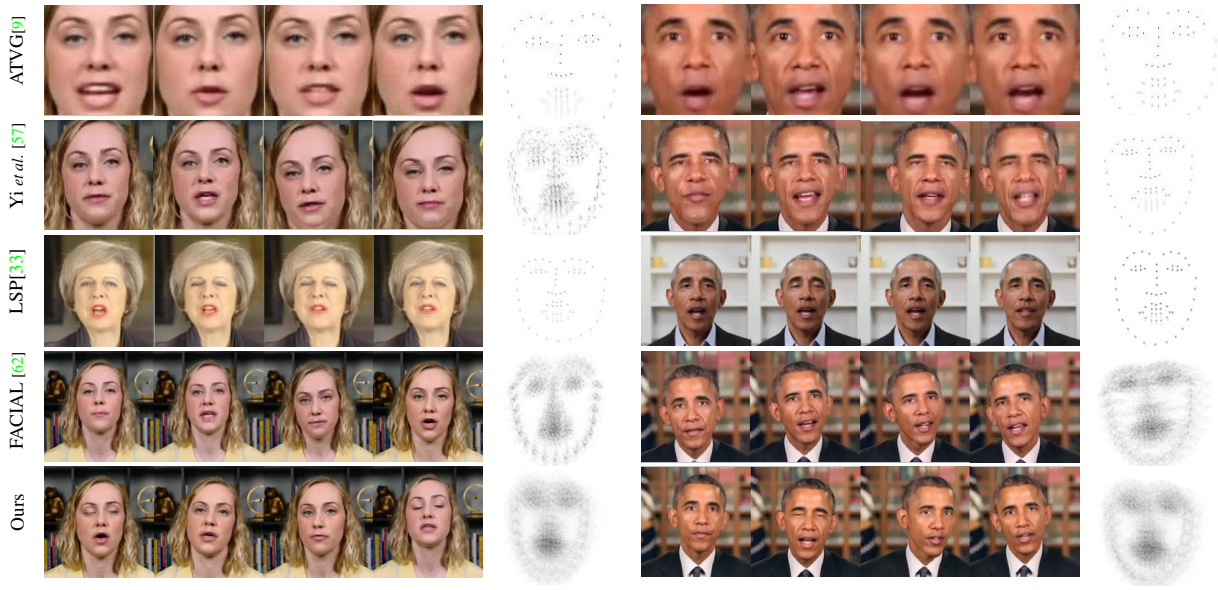


Figure 6. **Attention Weight Visualization.** The brighter white represents higher weights. The horizontal direction is along the time, and the vertical direction is along the feature dimension.

Effectiveness of Attention-based Modulator: Our Attention-based Modulator automatically assigns optimal attention weights on different features at each time step for each downstream generation task. It allows our model to extract as much useful information as possible from the entangled audio features for each specific downstream task and eliminate the interference sources. This is verified from the experimental results that our ATM variant outperforms Two-stream on all of the evaluation metrics summarized in Tab. 1.

Effectiveness of Pose Generation Network: Compared with simple MLP networks, the head pose dynamics generated by our PGN show superior perceptual results. The improvement comes from that our PGN decompose head pose sequence generation task into moving speed generation and moving direction generation. Firstly, the network is able to concentrate on the generation of moving speed which is more related to the rhythm of music, resulting in more rhythmical head pose dynamics that are in line with the music. This is verified in Tab. 1, that our method outperforms others a lot on the CCA metric of head pose. Then, the LSTM module for moving direction generation is able to consider not only the current audio features but also the generated head moving history, resulting in the more smooth and spontaneous head moving curves. As shown in Fig. 5, the visualization of pose rotation curves generated by our method (Ours) are smooth and resemble closely the ground truth. Specifically, the turn of the dominant vary-



a) Video frames b) Tracemaps c) Video frames d) Tracemaps

Figure 7. **Visual Comparison with State-of-the-art Methods.** a) and c) are the generated video frames. b) and d) are the corresponding tracemaps of facial landmarks across multiple frames. From the tracemaps, we can see our method generates the most diverse head pose dynamics.

ing angle (shown as green curve) of our generated head occurs nearly at the same time with ground truth. It's recommended to check our supplementary video to compare the difference more clearly.

Analysis of Attention Weight: To further investigate the role of the Attention Modulator (ATM), we visualize the predicted attention weights for synthesis tasks of facial expression, head movement, and eye state. As shown in the case illustrated in Fig. 6, it can be observed that:

- When there is background music only and no human voice, the ATM pays more emphasis on the stream of background music (melody feature), as shown in the earlier part of the music (See Region I).
- When there is both background music and human voice, in the tasks of face expression and head pose, the ATM modulates the weights between two streams to pursue more expressive results (See Region II). From the numerical viewpoint, the weights of the human voice are higher than that of background music. In this case, the human voice dominates the generation of face expression and head pose.
- When there is both background music and human voice, both the human voice and background music affect the long-time eye closing state, simultaneously (See Region III) or separately (See Regions IV and V).

4.4. Comparisons with State-of-the-art Methods

4.4.1 Compared State-of-the-art Methods

Most previous state-of-the-art methods are designed for talking scenarios and trained on talking datasets such as Voxceleb2 [10] and LRS2 [11]. For a fair comparison, we select and retrain the methods whose training code are open to the public on our SingingFace dataset. The selected compared state-of-the-art methods are as follows:

- ATVG [9] is one 2D-based cascade GAN approach to generate a talking face video that is robust to different facial characteristics, by taking an audio sequence and a target image as input.
- Yi *et al.* [57] utilize 3D face model information to synthesize photo-realistic talking face videos with personalized pose dynamics.
- LiveSpeechPortraits (LSP) [33] presents a live system utilizing 2D landmarks to generate personalized photorealistic talking-head animation in real time.
- FACIAL [62] integrates implicit attribute learning to synthesize 3D face animation with realistic motions of lips, head poses, and eye blinks.

We also report the qualitative comparison results with Song2Face [22], which is the only one method designed for singing scenarios up to now to the best of our knowledge. Song2Face maps each human voice segment to facial

expression and head rotation parameters, and uses an adaptive filter network to incorporate information from neighboring frames for temporal stability. It should be note that Song2Face only models with single driving source (plain human singing voice) as input, while ours supports multiple driving sources, *e.g.* human voices and background music, and focuses on how to collaborate with different driving sources to generate more realistic head movements. In addition, eye states are dealt with as a part of facial expression for Song2Face, while ours decompose the generation of eye states as an individual generation task. Since the implementation of Song2Face is unavailable, quantitative evaluation with Song2Face is absence in this paper. It's recommended to see our supplementary video for better comparison.

4.4.2 Qualitative Comparison

Fig. 7 and Fig. 8 shows the visual comparison with other state-of-the-art methods. We show the summary of qualitative comparison results in this section.

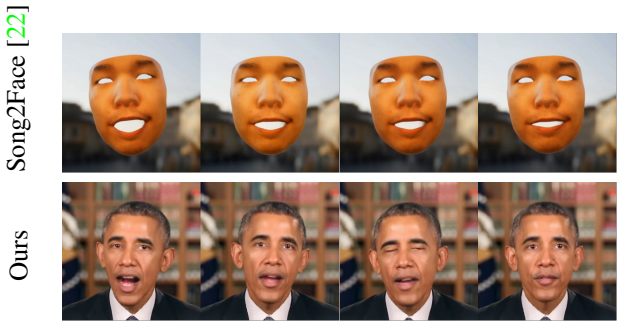


Figure 8. **Visual Comparison with Song2Face.** Our method can generate photo-realistic frames, diverse head pose and natural eye closing dynamics, which is infeasible for Song2Face [22].

Realism on Pose Dynamics: As shown in supplementary material, ATVG [9] only generates talking face videos with a fully static head pose, which is against the human common sense. Yi *et al.* [57] generate photo-realistic videos but the talking faces usually show subtle movements due to the supervision pipeline. In addition, the generated head pose dynamics behave discontinuously because of the background matching trick used in [57], which matches short-term generated head poses to one same target frame when the target frames are scarce in the target video. LiveSpeechPortraits [33] generates smooth but relatively small head movements. The generated head pose also shows a weak correlation with the rhythm of the music. FACIAL [62] and Song2Face [22] can generate more natural head pose dynamics than other compared state-of-art methods, but they still show only a few variations in head movement patterns over a long period of time. Our method can generate the most realistic head pose crediting to our pose generation method. To be specifically, for example, the head rotates quickly and dramatically during dense syllables, while slowly during pronouncing long syllables. Readers are recommended to watch the supplementary video to see the vivid visual results more clearly.

Realism on Eye States: The generation methods for eye states between the compared methods are various. ATVG and Yi *et al.* do not involve generating eye state parameters, therefore they do not produce any eye closing dynamics. For Song2Face and FACIAL, they learn random blink dynamics from data. However, Song2Face only performs well on plain human singing voice (no background music), and FACIAL only generates open eyes during inference, failing to generate spontaneous eye closing dynamics due to the complex entanglement between short random blinks and long-time eye closing states in SingingFace dataset. LiveSpeechPortraits directly sample random blink dynamics from target video and our method synthesizes random blinks from pre-defined random distributions, both of which show realistic random blink results. Moreover, as shown in Fig. 9, our method can also generate long-time eye closing dynamics (>0.5s) during voice based on the rhythm and emotion in the music, which further enhances the sense of realism.

4.4.3 Quantitative Evaluation

We use the same test set of music with the ablation study to compare our method with state-of-the-art counterparts. To clear out the effectiveness of the audio source separation model used in our method, we report the compared results on both mixed signals and separated signals (human voice and background music). Our method gets superior results on the most of metrics in the both cases. The results are summarized in Tab. 2.

Lip-sync metric: Similar to the ablation study, we evaluate the Landmark Distance Metric [8] and Audio-Visual Confidence score [12] to compare the lip synchronization of our method with the state-of-the-art methods. Tab. 2 shows that in both mixed and separated scenarios, our method beats all counterparts. It also shows that it is beneficial to separate the human voice from the plain music wave for mouth movement generation. Note that separated singing voice is given as input to the pre-trained lip-sync evaluation model during evaluation to ensure the pre-trained model performs correctly.

Pose Realism: In the evaluation of the realism of pose dynamics between different methods, we measure Canonical Correlation between predicted pose parameter sequences and ground-truth following with [64]. Similarly, to emphasize the evaluation of the rhythm of the synthesized head pose sequences, we apply Canonical Correlation Analysis to the speed of the head movement. Tab. 2 shows that our method generates more realistic and rhythmic pose

Table 2. **Comparisons with State-of-the-art Methods**

Method	Mixed Wave				Separated Wave				blinks/s	blink dur.(s)	CPBD↑
	AVC↑	LMD↓	CCA(pose)↑	CCA(eye)↑	AVC↑	LMD↓	CCA(pose)↑	CCA(eye)↑			
ATVG[9]	0.27	1.46	—	—	0.34	1.40	—	—	—	—	0.11
Yi *et al.*[57]	1.23	1.48	0.18	—	1.45	1.44	0.19	—	—	—	0.29
LiveSpeechPortraits[33]	0.31	1.43	**0.32**	—	0.35	1.42	0.32	—	—	—	0.20
FACIAL[62]	1.49	1.29	0.25	0.08	1.61	1.23	0.26	0.08	—	—	0.31
Ground Truth	3.00	—	—	—	3.00	—	—	—	0.35	0.23	0.37
Ours	**1.69**	**1.17**	0.26	**0.18**	**1.90**	**1.08**	**0.33**	**0.12**	0.38	0.26	**0.34**

Table 3. **Results of User Study.**

Method	Lip	Head	Eye	Conformity
ATVG[9]	1.24	1.24	1.20	1.19
Yi *et al.* [57]	2.13	1.54	1.51	1.70
LiveSpeechPortraits[33]	1.86	2.13	2.23	2.26
FACIAL[62]	3.51	3.53	2.71	3.56
Ours	**4.29**	**4.38**	**4.11**	**4.45**

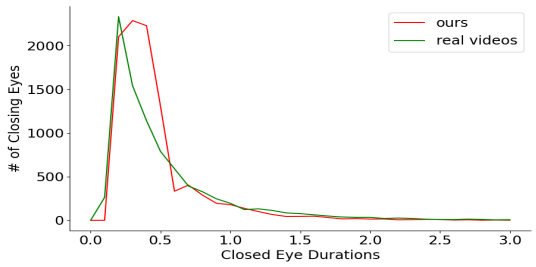


Figure 9. **Distribution of Eye Closing Duration.** Our method is able to generate realistic closed eye duration of similar distribution with the real videos.

dynamics.

Eye Realism: We measure Canonical Correlation between predicted eye state parameter sequences and ground-truth following with [64] to evaluate the realism of long-time eye closing dynamics. For random blinking evaluation, we count the average blinking frequency (blinks/s) and intra-blink duration (s) of generated singing face videos, and compare them with ground truth. As shown in Tab. 2, these two statistics of our method are similar to the ground truth, falling within a reasonable range.

Sharpness metric: Cumulative probability blur detection (CPBD) is evaluated to measure the generated frame sharpness of different methods. Our implementation of renderer module generates the most sharpness facial texture according to Tab. 2. However, as shown in our supplementary video, the generated texture of mouth region when opening mouth wide and the generated texture of eyelid when closing eyes look a little blur. The blur texture should come from the data scarcity of open mouth and closed eyes. It should be easy to improve the texture, simply by training the renderer with more data, or replacing the renderer with a few-shot face generator.

Tab. 2 shows the effectiveness of the audio source separation step for the singing face generation task, that almost all the evaluation metrics improve after decoupling human voice and background music. It also shows the superiority of our method, which generates the most realistic singing face videos and behaves better on all the evaluation metrics.

4.5. User Study

We invite 15 volunteers to evaluate our method and previous works. The volunteers are a group of college students with gender balance, no previous face synthesis study experience, but are informed of the study's purpose, the standard for evaluation, and the number of compared video groups before making evaluations. The volunteers are asked to make evaluations of videos group by group. In each group, 5 synthesized videos of compared methods are shown in order. There are 5 video groups in total. During evaluating each group, the volunteers were asked to watch all the videos of the group, then to score the videos at once based on the following criteria: 1) audio-visual synchronization, 2) natural head motion, 3) realistic eye state, 4) conformity with music. The evaluation scores include 1(very bad), 2(bad), 3(normal), 4(good), 5(very good). Finally we calculate the average scores for each method. As summarized in Tab. 3, our method has the superior visual realism than previous methods.

5. Discussion

Limitation and Future Work: The proposed method achieves more expressive results against previous methods. However, as shown in the supplementary video, under chaotic environments, our method fails like previous methods, which is because the adopted audio separator can not distinguish different human voices. On the other hand, this paper focuses on the synthesis of the head region, leaving the dynamics of the upper torso unsolved. We should note that it is more challenging to generate a realistic and expressive virtual human with dynamics of the upper torso and even the full body. This will be the direction of our future efforts. Moreover, our method just learns the implicit context from input audio, and it's indeed a interesting improvement direction to incorporate semantics from lyrics of songs.

Ethics Statement: The work itself does not uniquely raise any new ethical challenges. However, we must ac-

knowledge that the topic of image/video synthesis has been receiving many ethical concerns. These kinds of algorithms are vulnerable to malicious use, such as potentially misused to produce misleading information or violate the portrait right. Therefore, we appeal the research community and potential users to explore the techniques responsibly.

Acknowledgement

This work was supported in part by National Key R&D Program of China (Grant No. 2021YFC3300403), National Natural Science Foundation (Grant No. 62072382), and Yango Charitable Foundation. Guo was partially supported by National Science Foundation (OAC-2007661).

References

[1] O. Alemi, J. Françoise, and P. Pasquier. Groovenet: Real-time music-driven dance movement generation using artificial neural networks. *networks*, 8(17):26, 2017. 2

[2] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016. 7

[3] M. Brand. Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 21–28, 1999. 2

[4] C. Bregler, M. Covell, and M. Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 353–360, 1997. 1, 2

[5] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013. 7

[6] M. Cardle, L. Barthe, S. Brooks, and P. Robinson. Music-driven motion editing: Local motion transformations guided by music analysis. pages 38–44, 2002. 2

[7] L. Chen, G. Cui, C. Liu, Z. Li, Z. Kou, Y. Xu, and C. Xu. Talking-head generation with rhythmic head motion, 2020. 1, 2, 5

[8] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 520–535, 2018. 7, 10

[9] 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*, pages 7832–7841, 2019. 2, 3, 9, 10, 11

[10] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 9

[11] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 9

[12] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 7, 10

[13] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. Black. Capture, learning, and synthesis of 3D speaking styles. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10101–10111, 2019. 1, 2, 5

[14] D. Das, S. Biswas, S. Sinha, and B. Bhowmick. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, page 408–424, Berlin, Heidelberg, 2020. Springer-Verlag. 2

[15] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong. 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 (CVPR) Workshops*, June 2019. 7

[16] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. *ACM Transactions on Graphics (TOG)*, 21(3):388–398, 2002. 2

[17] O. Fried, A. Tewari, M. Zollhöfer, A. Finkelstein, E. Shechtman, D. B. Goldman, K. Genova, Z. Jin, C. Theobalt, and M. Agrawala. Text-based editing of talking-head video. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 2

[18] Y. Guo, K. Chen, S. Liang, Y.-J. Liu, H. Bao, and J. Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021. 2

[19] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50):2154, 2020. Deezer Research. 4, 7

[20] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 5

[21] R. Huang, H. Hu, W. Wu, K. Sawada, M. Zhang, and D. Jiang. Dance revolution: Long-term dance generation with music via curriculum learning. *arXiv preprint arXiv:2006.06119*, 2020. 2

[22] S. Iwase, T. Kato, S. Yamaguchi, T. Yukitaka, and S. Morishima. Song2face: Synthesizing singing facial animation from audio. In *SIGGRAPH Asia 2020 Technical Communications*, pages 1–4. 2020. 3, 9, 10

[23] 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*, pages 14080–14089, 2021. 1

[24] 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 (CVPR)*, pages 14080–14089, June 2021. 1

[25] H.-K. Kao and L. Su. Temporally guided music-to-body-movement generation. pages 147–155, 2020. 2

[26] H.-Y. Lee, X. Yang, M.-Y. Liu, T.-C. Wang, Y.-D. Lu, M.-H. Yang, and J. Kautz. Dancing to music. *arXiv preprint arXiv:1911.02001*, 2019. 2, 3

[27] J. Lee, S. Kim, and K. Lee. Listen to dance: Music-driven choreography generation using autoregressive encoder-decoder network. *arXiv preprint arXiv:1811.00818*, 2018. 2

[28] M. Lee, K. Lee, and J. Park. Music similarity-based approach to generating dance motion sequence. *Multimedia tools and applications*, 62(3):895–912, 2013. 2

[29] J. Li, Y. Yin, H. Chu, Y. Zhou, T. Wang, S. Fidler, and H. Li. Learning to generate diverse dance motions with transformer. *arXiv preprint arXiv:2008.08171*, 2020. 2

[30] L. Li, S. Wang, Z. Zhang, Y. Ding, Y. Zheng, X. Yu, and C. Fan. Write-a-speaker: Text-based emotional and rhythmic talking-head generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1911–1920, 2021. 2

[31] R. Li, S. Yang, D. A. Ross, and A. Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation. *arXiv preprint arXiv:2101.08779*, 2021. 2

[32] Y. Li, K. Swersky, and R. Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727. PMLR, 2015. 6

[33] Y. Lu, J. Chai, and X. Cao. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (TOG)*, 40(6):1–17, 2021. 9, 10, 11

[34] S. Marcos, J. Gómez-García-Bermejo, and E. Zalama. A realistic, virtual head for human–computer interaction. *Interacting with Computers*, 22(3):176–192, 2010. 1

[35] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009. 7

[36] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar. 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*, pages 484–492, 2020. 1, 2

[37] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 818–833, 2018. 1

[38] T. Shimba, R. Sakurai, H. Yamazoe, and J.-H. Lee. Talking heads synthesis from audio with deep neural networks. In *2015 IEEE/SICE International Symposium on System Integration (SII)*, pages 100–105. IEEE, 2015. 1, 2

[39] T. Shiratori, A. Nakazawa, and K. Ikeuchi. Dancing-to-music character animation. 25(3):449–458, 2006. 2

[40] S. Si, J. Wang, X. Qu, N. Cheng, W. Wei, X. Zhu, and J. Xiao. Speech2video: Cross-modal distillation for speech to video generation. *arXiv preprint arXiv:2107.04806*, 2021. 1

[41] S. Sinha, S. Biswas, and B. Bhowmick. Identity-preserving realistic talking face generation, 2020. 3, 6

[42] G. Sun, Y. Wong, Z. Cheng, M. S. Kankanhalli, W. Geng, and X. Li. Deepdance: music-to-dance motion choreography with adversarial learning. *IEEE Transactions on Multimedia*, 23:497–509, 2020. 2

[43] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 1, 2

[44] T. Tang, J. Jia, and H. Mao. Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis. pages 1598–1606, 2018. 2

[45] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *European Conference on Computer Vision*, pages 716–731. Springer, 2020. 1

[46] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *European Conference on Computer Vision*, pages 716–731. Springer, 2020. 1, 2

[47] J. Thies, S. Zollhofer, K. Marc, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. *arXiv preprint arXiv:2007.14808*, 2020. 1

[48] L. Wang, W. Han, F. K. Soong, and Q. Huo. Text driven 3d photo-realistic talking head. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011. 2

[49] S. Wang, L. Li, Y. Ding, C. Fan, and X. Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv preprint arXiv:2107.09293*, 2021. 2

[50] S. Wang, L. Li, Y. Ding, and X. Yu. One-shot talking face generation from single-speaker audio-visual correlation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2531–2539, 2022. 2

[51] Z. Wang, Z. Liu, Z. Chen, H. Hu, and S. Lian. A neural virtual anchor synthesizer based on seq2seq and gan models. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 233–236. IEEE, 2019. 1

[52] X. Wen, M. Wang, C. Richardt, Z.-Y. Chen, and S.-M. Hu. Photorealistic audio-driven video portraits. *IEEE Transactions on Visualization and Computer Graphics*, 26(12):3457–3466, 2020. 1

[53] T. Xie, L. Liao, C. Bi, B. Tang, X. Yin, J. Yang, M. Wang, J. Yao, Y. Zhang, and Z. Ma. Towards realistic visual dubbing with heterogeneous sources. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1739–1747, 2021. 2

[54] N. Yalta, S. Watanabe, K. Nakadai, and T. Ogata. Weakly-supervised deep recurrent neural networks for basic dance step generation. pages 1–8, 2019. 2

[55] X. Yao, O. Fried, K. Fatahalian, and M. Agrawala. Iterative text-based editing of talking-heads using neural retargeting. *ACM Transactions on Graphics (TOG)*, 40(3):1–14, 2021. 2

[56] Z. Ye, H. Wu, J. Jia, Y. Bu, W. Chen, F. Meng, and Y. Wang. Choreonet: Towards music to dance synthesis with choreographic action unit. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 744–752, 2020. 3

[57] R. Yi, Z. Ye, J. Zhang, H. Bao, and Y.-J. Liu. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137*, 2020. 1, 2, 3, 5, 9, 10, 11

[58] J. Yu and Z.-F. Wang. A video, text, and speech-driven realistic 3-d virtual head for human–machine interface. *IEEE transactions on cybernetics*, 45(5):991–1002, 2014. 1

[59] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9459–9468, 2019. 1

[60] C. Zhang, S. Ni, Z. Fan, H. Li, M. Zeng, M. Budagavi, and X. Guo. 3d talking face with personalized pose dynamics. *IEEE Transactions on Visualization and Computer Graphics*, 2021. 2

[61] C. Zhang, Y. Zhao, Y. Huang, M. Zeng, S. Ni, M. Budagavi, and X. Guo. Facial: Synthesizing dynamic talking face with implicit attribute learning. *arXiv preprint arXiv:2108.07938*, 2021. 1

[62] C. Zhang, Y. Zhao, Y. Huang, M. Zeng, S. Ni, M. Budagavi, and X. Guo. Facial: Synthesizing dynamic talking face with implicit attribute learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3867–3876, 2021. 2, 3, 5, 6, 7, 9, 10, 11

[63] Y. Zhang, S. Zhang, Y. He, C. Li, C. C. Loy, and Z. Liu. One-shot face reenactment. *arXiv preprint arXiv:1908.03251*, 2019. 1

[64] Z. Zhang, L. Li, Y. Ding, and C. Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 7, 10, 11

[65] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9299–9306, 2019. 2

[66] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4176–4186, 2021. 1, 3

[67] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020. 3

[68] W. Zhuang, Y. Wang, J. Robinson, C. Wang, M. Shao, Y. Fu, and S. Xia. Towards 3d dance motion synthesis and control. *arXiv preprint arXiv:2006.05743*, 2020. 2