EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Multimodal-Driven Emotion-Controlled Facial Animation Generation Model

Zhenyu Qiu[1], Yuting Luo[2]*, Yiren Zhou[3], Teng Gao[1]

1. School of Computer and Information Engineering, Nanchang Institute of Technology, Nanchang, 330044, China
2. School of Music, Nanchang Institute of Technology, Nanchang, 330044, China
3. Rapid prototyping institute, Nanchang Institute of Technology, Nanchang, 330044, China

Abstract

INTRODUCTION: In recent years, the generation of facial animation technology has emerged as a prominent area of focus within computer vision, achieving varying degrees of progress in lip-synchronization quality and emotion control.
OBJECTIVES: However, existing research often compromises lip movements during facial expression generation, thereby diminishing lip synchronisation accuracy. This study proposes a multimodal, emotion-controlled facial animation generation model to address this challenge.
METHODS: The proposed model comprises two custom deep-learning networks arranged sequentially. By inputting an expressionless target portrait image, the model generates high-quality, lip-synchronized, and emotion-controlled facial videos driven by three modalities: audio, text, and emotional portrait images.
RESULTS: In this framework, text features serve a critical supplementary function in predicting lip movements from audio input, thereby enhancing lip-synchronization quality.
CONCLUSION: Experimental findings indicate that the proposed model achieves a reduction in lip feature coordinate distance (L-LD) of 5.93% and 33.52% compared to established facial animation generation methods, such as MakeItTalk and the Emotion-Aware Motion Model (EAMM), and a decrease in facial feature coordinate distance (F-LD) of 7.00% and 8.79%. These results substantiate the efficacy of the proposed model in generating high-quality, lip-synchronized, and emotion-controlled facial animations.

Keywords: Deep Learning; Computer Vision; Generative Adversarial Networks; Facial Animation Generation Technology; Multimodal

Received on 21 October 2025, accepted on 07 July 2025, published on 17 July 2025

doi: 10.4108/eetsis.7624

* Corresponding author. Email: 19979129968@163.com, yuting_luo24@outlook.com

1. Introduction

With the continuous advancement of generative models [1-4], facial animation generation technology [5-9] has rapidly developed and gained widespread attention. It is now applied in virtual character creation [10-12], animation film assistance, and short video production [13]. Current facial animation generation technologies can be categorized based on their driven targets into two types: (1) methods based on portrait videos [14-18], where the driven target is a portrait video, typically edited and reshaped directly or indirectly, and (2) methods based on portrait images [19-23], where the driven target is usually a single-frame portrait image, aiming to synthesize corresponding video segments through driving factors.

In high-quality facial animation technology, synchronizing speech content with lip movements is crucial. In contrast, high-fidelity visuals and rich facial expressions can further enhance the authenticity and practicality of the animation [5]. Both types of facial

animation generation technologies have explored facial emotional expression. Ji et al. [14] and Fang et al. [22] attempted to control facial emotions in videos without adding extra driving factors. Still, their methods were influenced by a single driving source, resulting in poor flexibility and low robustness. Other works [15, 23-24] introduced one or more additional driving factors to control facial emotions in videos. Although these methods granted emotional expression to the videos, the driving factors' complexity reduced lip synchronisation accuracy, interfering with the portrait background and limiting practical application.

This paper aims to design a highly flexible facial animation generation model that ensures precise lip movement and controllable emotions to improve the practical efficiency of facial animation generation technology. However, achieving such a model faces the following challenges: 1) ensuring instant usability while maintaining the identity of the target portrait and lip synchronization, avoiding unnecessary and cumbersome preprocessing; 2) managing variations in portrait facial expressions, which can easily cause non-rigid facial structure distortions that interfere with lip movements, reducing the accuracy of lip synchronization; 3) addressing the limitations of a single driving factor, which offer weak control, while multiple driving factors can easily lead to entanglement, resulting in the mixing of portrait contours and background pixels, and even causing ghosting and distortion in the visuals.

This paper proposes a multimodal-driven, emotion-controlled facial animation generation model to address these challenges. The model takes an expressionless static portrait as the target and uses audio, text, and emotional portrait images as driving factors. Building on traditional audio-driven facial animation generation techniques, it controls the facial expressions of the target portrait using emotional portrait images. It utilizes text features to assist in predicting lip movement changes based on audio, thereby generating high-quality lip-synchronized and emotion-controlled facial videos.

The network structure of this method comprises two end-to-end modules: the facial feature coordinate generation module and the facial video generation module, both of which are custom encoder-decoder deep learning networks. Additionally, to link these modules, facial feature coordinates are used to avoid direct reconstruction of target portrait pixels by the driving factors and prevent pixel mixing caused by multiple driving factors. Using three encoders, the facial feature coordinate generation module extracts different feature vectors from four inputs. Then, it predicts a sequence of facial coordinate offsets corresponding to the audio length through a decoder. After adding this to the target portrait feature coordinates, a new sequence of facial feature coordinates is generated. The facial video generation module uses the generated sequence of facial feature coordinates as driving factors to control the variation of portrait pixels, resulting in lip-synchronized and emotion-controlled facial animations. The qualitative and quantitative experiments and ablation studies conducted in this paper validate the necessity of the sub-network modules and the overall effectiveness of the model.

The main contributions of this paper are as follows:

A novel technical model is designed to generate high-quality, lip-synchronized and emotion-controlled facial animations driven by three modalities of data based on any expressionless target portrait.

A method is proposed to assist audio-driven inputs with text features, reducing distortions caused by expression changes in the lip region and thereby improving the accuracy of lip synchronization in generated facial animations.

The serial structure between modules ensures instant usability while preventing interference between multiple driving factors, guaranteeing the authenticity of the generated animations.

The paper is structured as follows: it reviews existing methods in the Related Work section, followed by a detailed Model Description of the proposed multimodal approach. The paper then presents Experiments and Results Analysis, highlighting qualitative and quantitative results, and concludes with the Conclusions and future work.

2. Related Works

2.1 Facial Animation Generation Technology

With the development of internet technology, the demand for high-quality facial animation videos is constantly rising. This trend has significantly influenced computer vision, computer generation, and computer graphics, prompting many technology companies and research institutions to use computer and artificial intelligence technologies to enhance video content quality, deepen creation automation, and reduce labour costs. Facial animation generation technology is particularly prominent among these [25-29].

Over the past five years, the emergence of excellent audiovisual databases [30-32] has accelerated the development of facial animation generation technology. Suwajanakorn et al. [32] synthesized lip movements from audio based on Obama's speech video and combined them with the original portrait data. Although the video results were impressive, the method was limited to specific identities and required a long training cycle. In the same

year, Chung et al. [33] proposed an encoder-decoder network based on Convolutional Neural Networks (CNN) [34], advancing research in generating facial animations during the image-to-image translation process. Following this, Chen et al. [19] proposed a concatenated structure of audio transformation and visual generation network (ATVGnet), which introduced facial feature coordinates as an intermediate representation, generating facial animation for any portrait while ensuring lip synchronization. Zhou et al. [21] further enhanced the realism of facial animations by adding a spontaneous head movement module. Additionally, Yu et al. [35] attempted to incorporate text as a driving factor on top of existing technologies to achieve multimodal complementarity, improving lip synchronization in videos. While these methods focus on the flexibility of video generation and the accuracy of lip synchronization, the input target constrains facial expressions, and emotional expression remains relatively singular, limiting practical application scenarios.

2.2 Emotion-Controlled Facial Animation Generation Technology

Facial expressions are one of the important ways humans express emotions. Richly animated facial expressions can enhance the appeal of human-computer interaction interfaces and virtual characters, making them more engaging for users. Therefore, most facial animation generation technologies proposed in recent years focus on generating facial expressions while ensuring lip synchronization, and can be categorized into two types based on the driven targets.

Emotion-Controlled Facial Animation Generation Technology Based on Portrait Videos: The goal is to reshape lip movements and edit facial expressions in the target video through driving factors. Ji et al. [14] used cross-reconstruction emotional decomposition techniques to decouple content features and emotional features from the original audio, combining them with portrait features to generate emotion-controlled facial feature coordinates. Finally, they produced portrait videos using target-adaptive face synthesis technology. Although this method addresses the issue of limited facial expression variety, its reliance on target-specific training reduces flexibility. Recently, Liang et al. [15] proposed a Granularly Controlled Audio-Visual Talking Heads (GC-AVT) model that comprehensively considers fine-grained control of lip movements, head poses, and facial expressions in facial animations. This model takes four videos as input and uses Generative Adversarial Networks (GAN) [36] and masking techniques [37] to reconstruct target videos, controlling lip movements, head poses, and facial expressions. However, the singularity of input modalities and the abundance of driving factors result in poor model robustness and overly random facial animation realism. Additionally, this method has limited capability in handling portraits and backgrounds, leading to simplistic scenes and low-resolution issues.

Emotion-Controlled Facial Animation Generation Technology Based on Portrait Images: This approach aims to synthesize lip-synchronized and emotion-controlled video segments frame by frame through driving factors. Compared to video-driven methods, image-driven methods require only one image to complete the generation task, offering greater flexibility. However, this flexibility comes at the cost of reduced controllability due to the limited input information.

Fang et al. [22] used a custom GAN model to directly drive audio synthesis pixel points. While generating facial expressions, audio noise also affected the process, severely reducing image fidelity. To enhance control over facial expressions, Eskimez et al. [23] and Agarwal et al. [38] introduced extra driving factors to control facial expressions coarse-grained. Eskimez et al. [23] added emotional labels, while Agarwal et al. [38] used emotional portraits. However, both methods share the common issue of dependence on training, and GANs are notoriously difficult to train, with poor stability, limiting their practical application based on target identities. Ji et al. [24] proposed the Emotion-Aware Motion Model (EAMM), which addresses the dependency on training in this direction. The model controls the identity features, lip movements, facial emotions, and head poses of generated videos through four inputs. However, due to the interplay of multiple driving factors, the overall control difficulty of the model is high, leading to ghosting and distortion in portraits, and the accuracy of the generated lip movements is relatively low.

In summary, the technology for generating emotion-controllable facial animations is still exploratory. Current methods struggle to cover all aspects, with significant room for improvement in model flexibility, lip-sync accuracy, richness of facial expressions, and diversity of head poses. This paper adopts a technology route based on portrait images to address these issues and introduces additional driving factors to control facial expressions. Unlike [23, 38], this paper requires additional textual information as input to provide context and content for assisting audio predictions of lip changes. Compared to the multi-target control of [24], this study abandons the issue of head pose diversity and focuses on model flexibility, lip-sync accuracy, and richness of facial expressions. The aim is to control facial expressions while minimizing ghosting and distortion in the visuals, thereby improving lip-sync accuracy. Furthermore, the network framework in this paper differs from the methods above by introducing facial coordinates as an intermediate representation, dividing the

model into two subnetwork modules. This approach avoids the direct driving of pixel points, enhancing the robustness of the generative model.

3. Model Description

This paper proposes a multimodal-driven, emotion-controllable facial animation generation model that focuses on the impact of facial expression reshaping on lip movements, as shown in Figure 1. The model utilizes four types of input data, with the target object being a neutral portrait image (located at the bottom left of the framework), while the other inputs act as driving factors. The driving factors include audio, text, and emotional portraits. Audio and text driving factors predict and generate lip movements, while the emotional picture is used to reshape facial expressions. Notably, a single audio signal may lack clarity or be affected by noise interference, and this model is additionally affected by facial expression reshaping, significantly increasing the difficulty of lip-sync compared to traditional methods. This paper provides context and content understanding through textual information to address these issues, helping the audio driving factor more accurately capture contextual lip changes to achieve multimodal fusion. The text features greatly enhance the accuracy of lip-sync predictions, avoiding distortions caused by facial expression reshaping and improving the realism of the generated videos.

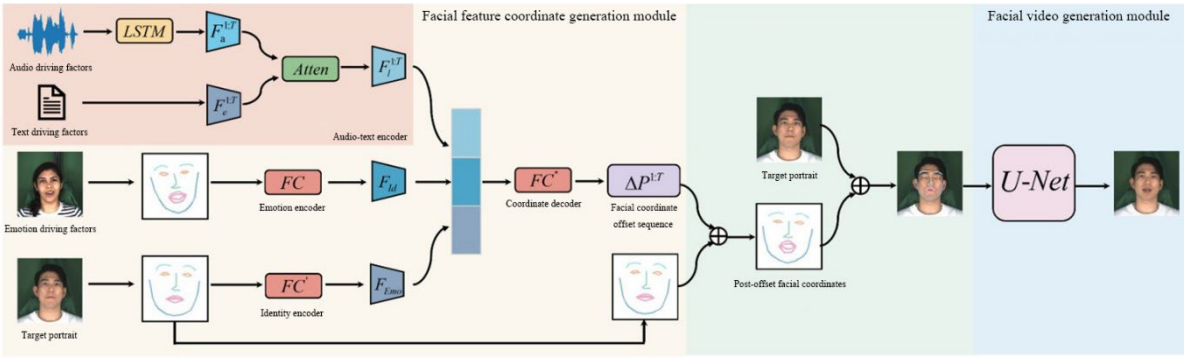

Fig. 1 Framework of the proposed model

The model is divided into two key modules: the facial feature coordinate generation module and the facial video generation module. The facial feature coordinate generation module generates a sequence of lip-sync and emotion-controllable facial coordinate offsets, which serve as intermediate features for the model. The facial video generation module then converts this predicted feature coordinate offset sequence into a facial video of the target portrait. A transformation process between the two modules is required to ensure seamless integration. The brief formulas for these modules are as follows:

$$\Delta \boldsymbol{P}^{1:T} = Mul2Lm(\boldsymbol{A}^{1:T}, \boldsymbol{E}, \boldsymbol{P}_{\text{Emo}}, \boldsymbol{P}_{\text{Id}}) \quad (1)$$

$$\boldsymbol{Q}^{1:T} = \Delta \boldsymbol{P}^{1:T} + \boldsymbol{P}_{\text{Id}} \quad (2)$$

$$\boldsymbol{V}^{1:T} = Lm2Vid(\boldsymbol{Q}^{1:T}, \boldsymbol{I}) \quad (3)$$

Where: $Mul2Lm$ denotes the facial feature coordinate generation module; $Lm2Vid$ denotes the facial video generation module; $\boldsymbol{A}^{1:T}$ represents the audio driving factor; $\boldsymbol{E}$ denotes the text driving factor; $\boldsymbol{P}_{\text{Emo}}$ represents the facial feature coordinates of the emotional driving factor; $\boldsymbol{P}_{\text{Id}}$ represents the facial feature coordinates of the target portrait; T denotes the length of the audio; $\Delta \boldsymbol{P}^{1:T}$ represents the predicted facial coordinate offset sequence; $\boldsymbol{Q}^{1:T}$ denotes the facial feature coordinate sequence of the target portrait after the offset; $\boldsymbol{I}$ represents the original image of the target portrait; $\boldsymbol{V}^{1:T}$ denotes the final generated video.

3.1 Facial Feature Coordinate Generation Module

The facial feature coordinate generation module adopts a GAN framework consisting of a generator and a discriminator. The generator consists of three encoders and one decoder: the audio-text encoder, the identity encoder, the emotion encoder, and the facial feature coordinate decoder. The model also includes two discriminators: one discriminating the authenticity of the generated coordinates (the authenticity discriminator) and another discriminating the image frame rate (the image frame rate discriminator).

3.1.1 Data Preprocessing

Before being used as input, the three modalities of data undergo different preprocessing steps. Specifically, the portrait images need to utilize a pre-trained facial feature keypoint recognition algorithm [39] to obtain facial feature coordinates $\boldsymbol{P}_{\text{Id}}$ $\boldsymbol{P}_{\text{Emo}}$ with a dimension of 68×3, reducing the complex pixel points into a unified representation. Additionally, the detected coordinates need to be normalised to minimize the differences between different target portrait coordinates. The audio data is processed using Python's Libros package, converting the audio driving factor $\boldsymbol{A}^{1:T}$ into Mel Frequency Cepstral Coefficients (MFCC) $\boldsymbol{A}_{\text{MFCC}}^{1:T}$ with a dimension of (T ,40) to suit the subsequent network structure better and enhance lip-sync accuracy. The text data requires using the open-source deep learning toolkit Merlin to extract text feature vectors with a dimension of (T , 425).

3.1.2 Encoder Network Structure

The audio-text encoder aims to extract multimodal latent features that drive lip changes from audio and text features, thereby predicting the changes in facial coordinate sequences. This encoder consists of two network models: a Long Short-Term Memory (LSTM) network [40] and a Self-Attention mechanism [41]. The preprocessed audio driving factor's Mel Frequency Cepstral Coefficients $\boldsymbol{A}_{\text{MFCC}}^{1:T}$ are input into the LSTM to extract audio feature vectors $\boldsymbol{F}_{a}^{1:T}$ with a dimension of (T , 256). To improve the accuracy of lip movement predictions and enhance the overall effectiveness of the driving factors, this paper employs the encoder structure of the Transformer model [41]. The goal is to capture long-term dependencies between the audio feature vector $\boldsymbol{F}_{a}^{1:T}$ and the text feature vector $\boldsymbol{F}_{e}^{1:T}$ through Self-Attention while improving the model's overall performance through the complementarity of different modalities. Specifically, after linearly merging the audio feature vector $\boldsymbol{F}_{a}^{1:T}$ and the text feature vector $\boldsymbol{F}_{e}^{1:T}$, the result is input into the Self-Attention mechanism encoder $Atten$ to extract common features, resulting in latent audio-text features $\boldsymbol{F}_{l}^{1:T}$ with a dimension of (T 256). The specific formulas for the audio-text encoder are as follows:

$$\boldsymbol{F}_{l}^{1:T} = Atten(LSTM(\boldsymbol{A}_{\text{MFCC}}^{1:T}), \boldsymbol{F}_{e}^{1:T}) \tag{4}$$

The identity encoder FC takes the facial feature coordinates $\boldsymbol{P}_{\text{Id}}$ of the target portrait as input, aiming to capture and extract the identity features $\boldsymbol{F}_{\text{Id}}$ of the target portrait. This encoder uses a fully connected network with five layers, with dimensions of (204, 256), (256, 256), (256, 256), (256, 256), and (256, 256). A Leaky Rectified Linear Unit (LeakyReLU) with a slope of 0.2 is used as the activation function after each fully connected layer to introduce non-linearity. The specific formulas for the identity encoder are as follows.

$$\boldsymbol{F}_{\text{Id}} = FC(\boldsymbol{P}_{\text{Id}}) \tag{5}$$

The emotion encoder FC' takes the facial feature coordinates $\boldsymbol{P}_{\text{Emo}}$ of the emotional portrait as input, aiming to extract the emotional features $\boldsymbol{F}_{\text{Emo}}$ of the emotional picture and thereby reshape the facial expressions of the target portrait. Similar to the identity encoder, this network also uses a fully connected network with five layers. The specific formulas for the emotion encoder are as follows:

$$\boldsymbol{F}_{\text{Emo}} = FC'(\boldsymbol{P}_{\text{Emo}}) \tag{6}$$

3.1.3 Decoder Network Structure

The facial feature coordinate decoder FC'' takes the latent audio-text features $F_{l}^{1:T}$, identity features $\boldsymbol{F}_{\text{Id}}$, and emotional features $\boldsymbol{F}_{\text{Emo}}$ as input, aiming to predict the facial coordinates aligned with the audio text through the latent audio-text features and reshape the facial expression coordinates of the target portrait using the emotional features. [42] develops a multimodal intelligent model to recognize four specific emotions in drivers, which can impact driving performance. By analyzing motor activity signals and facial geometry images using machine learning, the model achieves 96% accuracy in a simulated environment. It combines data from driver behaviour and facial features, processed through a CNN, to accurately classify emotions, confirming a significant link between motor activity, behaviour, facial geometry, and induced emotions.

First, the identity and emotional features are concatenated frame by frame with the latent audio-text features to obtain total latent features with a dimension of (T , 768). The total latent features are then fed frame by frame into a fully connected network with five layers, with dimensions of (768, 512), (512, 256), (256, 256), (256, 256), and (256, 204), to obtain the facial coordinate offset sequence $\Delta \boldsymbol{P}^{1:T}$. Adding this $\Delta \boldsymbol{P}^{1:T}$ to the target portrait $\boldsymbol{P}_{\text{Id}}$ results in a lip-synchronized and emotion-controllable

facial feature coordinate sequence $\boldsymbol{Q}^{1:T}$. The specific formulas for the facial feature coordinate decoder are as follows.

$$\boldsymbol{Q}^{1:T} = FC''(\boldsymbol{F}_l^{1:T}, \boldsymbol{F}_{\mathrm{Id}}, \boldsymbol{F}_{\mathrm{Emo}}) + \boldsymbol{P}_{\mathrm{Id}} \tag{7}$$

3.1.4 Discriminator Network Structure

The authenticity discriminator follows the design philosophy of GANs, aiming to discern the authenticity of the generated facial coordinates. This discriminator employs a five-layer fully connected neural network, with dimensions of (204, 256), (256, 256), (256, 128), (128, 64), and (64, 1) for each layer. Except for the last layer, each layer uses a Leaky ReLU activation function with a slope of 0.2 to rectify the linear units. Additionally, to optimize the discriminator further, this paper sets a loss function for the discriminator to impose further constraints on the model. The specific formula is as follows:

$$L_{\mathrm{DR}} = log\, D_R(\hat{\boldsymbol{Q}}^t) + log(1 - D_R(\boldsymbol{Q}^t)) \tag{8}$$

Where $\boldsymbol{Q}^t$ represents the predicted frame facial feature coordinates $\hat{\boldsymbol{Q}}^t$, the reference frame facial feature coordinates D_R, the authenticity discriminator, and L_{DR} the authenticity discriminator loss function.

The image frame rate discriminator aims to enhance the coherence and smoothness of the predicted facial feature coordinate sequence, preventing excessive jitter between adjacent frames. This discriminator consists of three three-layer fully connected networks. The first two networks input the facial coordinates of corresponding frames, capturing their feature vectors with dimensions of (204, 256), (256, 128), and (128, 64) for each layer. The extracted features are then concatenated and input into the third network, capturing the correlation between them, with dimensions of (128, 64), (64, 32), and (32, 1) for each layer. Similar to the authenticity discriminator, this discriminator utilizes the Leaky ReLU activation function. The specific formula for the discriminator loss function is as follows:

$$L_{DF} = \log D_F\left(\hat{\boldsymbol{Q}}^{t+1}, \hat{\boldsymbol{Q}}^t\right) + \log\left(1 - D_F\left(\hat{\boldsymbol{Q}}^{t+1}, \boldsymbol{Q}^t\right)\right) \tag{9}$$

Where $\hat{\boldsymbol{Q}}^{t+1}$ represents the $t+1$ reference frame facial feature coordinates D_F, the image frame rate discriminator, and L_{DF} the image frame rate discriminator loss function.

3.1.5 Loss Functions

To further optimize the quality of the generated feature coordinates, this paper adds two additional loss functions to the authenticity discriminator loss function L_{DR} and the image frame rate discriminator loss function L_{DF}, focusing on the generation effects of the model's lip coordinates. These include a facial coordinate loss function L_F aimed at improving the facial expression in the image and a lip coordinate loss function L_L aimed at optimizing lip synchronization.

Specifically, the facial coordinate loss function L_F is designed to enhance the accuracy and authenticity of facial coordinates by minimizing the Euclidean distance between the predicted coordinates and the reference coordinates, which is a common approach in this area of research. The specific calculation formula is as follows:

$$L_F = \sum_{t=1}^{T}\sum_{i=1}^{N}\left\|Q^{t,i} - \hat{Q}^{t,i}\right\|_2^2 \tag{10}$$

Where $Q^{t,i}$ represents the predicted facial feature coordinates at index i for the frame; $\hat{Q}^{t,i}$ denotes the reference facial feature coordinates at index i for the frame; T means the length of the audio; and N represents the total number of index points for the facial coordinates.

The changes in the lip region coordinates will be more complex to align and match the audio and text driving factors. To address this, this paper adds a lip coordinate loss function L_L. This function ensures the detail of the changes in the lip region by calculating the Laplacian distance between the predicted coordinates and the reference coordinates [43], thus improving the lip synchronization effect. The calculation formula is as follows:

$$L_L = \sum_{t=1}^{T}\sum_{i=1}^{M}\left\|\mathrm{L}(Q_{lip}^{t,i}) - \mathrm{L}(\hat{Q}_{lip}^{t,i})\right\|_2^2 \tag{11}$$

Where $Q_{lip}^{t,i}$ represents the predicted lip feature coordinates at index i for the frame; $\hat{Q}_{lip}^{t,i}$ denotes the reference lip feature coordinates at index i for the frame; L indicates the Laplacian distance; and M represents the total number of index points for the lip coordinates, which is 20.

The authenticity discriminator loss function L_{DR} and the image frame rate discriminator loss function L_{DF} have been previously introduced. The overall loss function formula for the model is as follows:

$$L_{M2L} = \alpha L_F + \alpha L_L + \beta L_{DR} + \beta L_{DF} \tag{12}$$

Where L_{M2L} represents the total loss function for the facial feature coordinate generation module, and the weights for each sub-loss function are denoted as $\{\alpha = 0.1, \beta = 0.01\}$

3.2 Facial Video Generation Module

The facial video generation module utilizes the concept of image translation, aiming to transform the input images into target images using an end-to-end algorithm. This module comprises both a generator and a discriminator. The generator adopts a U-Net structure [44], with skip connections and a U-shaped convolutional neural network. The U-Net can capture local detail features and overall facial characteristics from the portrait images, offering advantages such as efficient training, high speed, and strong scalability. The discriminator serves as an image quality discriminator to enhance the quality of the generated images.

3.2.1 Data Preprocessing

In this paper, the predicted facial feature coordinate sequences $\boldsymbol{Q}^{1:T}$ are converted into image data, specifically by sequentially connecting the indices of the facial coordinates for each frame by region and using predefined colours to draw a sequence of RGB images with dimensions $(T\,256,\ 256,\ 3)$. Subsequently, the target portrait $\boldsymbol{I}$, also a three-channel image, is added frame by frame to this image sequence, resulting in a six-channel image sequence with dimensions $(T\,256,\ 256,\ 6)$.

3.2.2 Generator Network Structure

The generator adopts a traditional nine-layer U-Net structure comprising four encoder layers, one middle layer, and four decoder layers. The encoder layers consist of convolutional and pooling layers that progressively reduce the size of the feature maps while extracting high-level features from the images. The middle layer consists of multiple convolutional layers, aiming to maintain the size of the feature maps and extract richer features. The decoder layers consist of transposed convolutional layers and convolutional layers that gradually increase the size of the feature maps to generate outputs that match the dimensions of the input images. Furthermore, skip connections between the encoder and decoder layers link low-level and high-level features for feature sharing and fusion. The specific formula is shown in (3).

3.2.3 Discriminator Network Structure

The image quality discriminator aims to improve the image quality of the generated videos and maintain the consistency of the target portrait's identity. This discriminator is composed of a five-layer 2D-CNN, with the output channel numbers 64, 128, 256, 512, and 512 for the five layers, respectively. The convolutional kernel has a stride of 2 and a size of 3×3. The outputs are then flattened and fed into a two-layer fully connected network to distinguish the authenticity of the frame. Similarly, except for the last layer, each layer employs a LeakyReLU activation function with a slope of 0.2. The discriminator loss function is as follows.

$$L_{DQ} = \log D_Q\left(\hat{V}^t\right) + \log\left(1 - D_Q\left(V^t\right)\right) \tag{13}$$

where V^t represents the t frame image of the generated video $\hat{V}^t$, the t frame image of the reference video D_Q, the image quality discriminator, and L_{DQ} the image quality discriminator loss function.

3.2.4 Loss Function

This module, based on the image quality discriminator loss function L_{DQ}, additionally incorporates a perceptual loss function L_{VGG} to enhance the quality of the generated images. The perceptual loss function L_{VGG} utilizes a pre-trained 19-layer convolutional neural network [45], VGG-19 (Visual Geometry Group 19), to extract features from the generated and reference videos and compute the distance between these features. This module's perceptual loss function formula and the total loss function formula are shown as follows.

$$L_{VGG} = \left\| \theta\left(V_t\right) - \theta\left(\bar{V}_t\right) \right\|_1 \tag{14}$$

$$L_{L2V} = L_{VGG} + \lambda L_{DQ} \tag{15}$$

where θ represents the pre-trained VGG-19 network L_{VGG}, the perceptual loss function L_{L2V}, the total loss function of the facial video generation module, and $\lambda = 0.01$ the weight of the loss functions.

4. Experiments And Results Analysis

4.1 Experimental Setup

The method presented in this paper employs a serial structure, with the two network modules trained separately. Due to the differing designs and purposes of the modules, this study utilized the Multi-view Emotional Audiovisual Dataset (MEAD) [31] and VoxCeleb2 [30] audiovisual databases for training the network modules. The facial feature coordinate generation module utilized the MEAD [31] database, a high-resolution emotional audiovisual dataset featuring 60 participants from different continents and encompassing 8 types of facial emotions and corresponding audio-text materials. A subset of 43 frontal poses of portraits from this database was used for training, testing, and validation of the model. The facial video

generation module utilized the VoxCeleb2 [30] audiovisual database, a diverse identity audiovisual dataset. The original videos were downloaded via URLs provided by VoxCeleb2, collecting approximately 2,000 identity sources for the model training, testing, and validation.

The framework was implemented using PyTorch 1.8.0 and CUDA 11.1. Four Nvidia Tesla V100S GPUs, each equipped with 32GB of memory, were used to train both modules. The model's learning rate is set to 0.0001, and the total training duration is approximately 24 hours.

4.2 Qualitative Experiments

To effectively demonstrate the generation quality of the model, this study designed qualitative experiments and compared the results with state-of-the-art methods. The selected comparison methods include ATVGnet [19], MakeItTalk [21], and EAMM [24]. Two target portraits were randomly chosen from the audiovisual databases MEAD [31] and VoxCeleb2 [30]. Notably, both ATVGnet [19] and MakeItTalk [21] require only the target portrait and audio driving factors as input, while the method proposed in this paper additionally requires an emotional picture and text driving factors. The emotional portraits used in this qualitative experiment were all selected from the happy emotion portraits in MEAD [31]. Furthermore, EAMM [24] required input of the target portrait, head pose sequences, audio driving factors, and video driving factors; the additional head pose sequences and video driving factors were also sourced from MEAD [31].

The results of the qualitative experiments are shown in Figure 2, where GT represents the ground truth data. ATVGnet [19] produces relatively realistic lip shapes. However, the generated visuals are limited to the facial subject, resulting in somewhat stiff expressions. MakeItTalk [21] could control slight head rotations to some extent, and the lip movements are fairly aligned, but the facial expressions remain constrained by the input target portrait. EAMM [24] considered lip movements, head poses, and facial expressions, but its strict input requirements resulted in less stable generation effects and average practical usability. In contrast to the methods above, the animations generated by the process in this paper exhibit vivid facial expressions, accurate mouth shapes, and higher lip-sync quality, showing less discrepancy from the real data. This experiment demonstrated the effectiveness of the multimodal approach proposed in this paper.

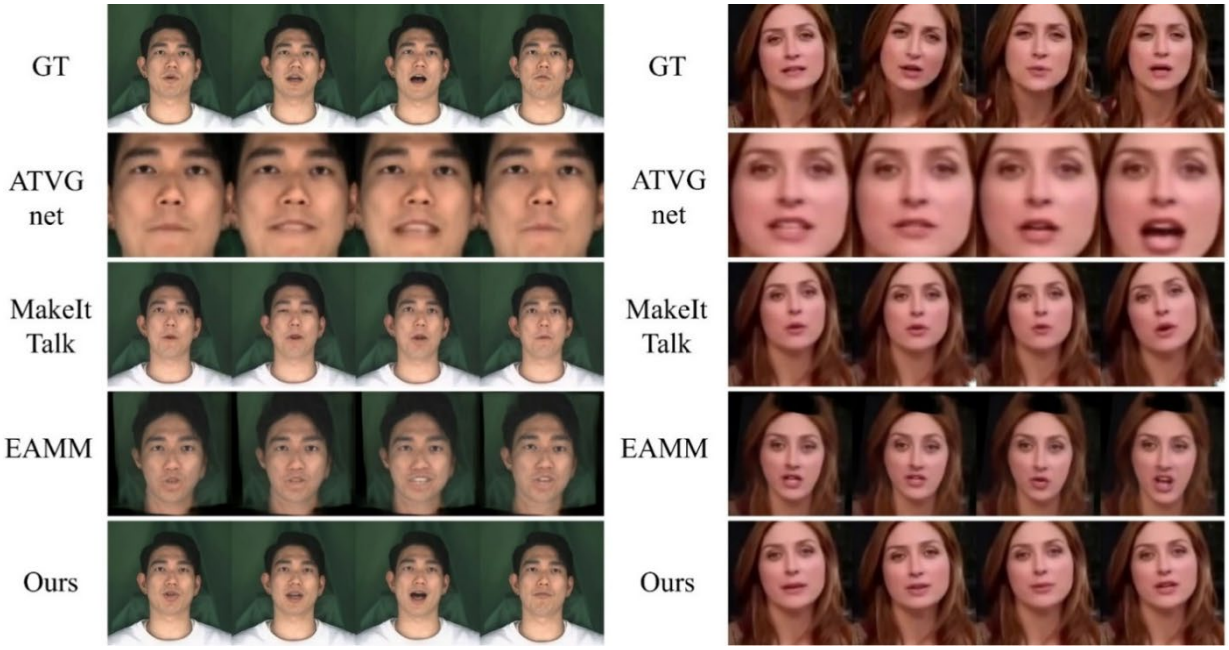

Fig. 2 Results of Qualitative Experiments

4.3 Quantitative Experiments

The quantitative experiments compared methods based on ATVGnet [19], MakeItTalk [21], EAMM [24], and real data while also including baseline methods from the MEAD [31] audiovisual database. The experimental subjects were selected from both MEAD [31] and VoxCeleb2 [30] audiovisual databases, utilizing evaluation metrics. The Lip Landmarks Distances (L-LD) were used to assess the synchronization of lip shapes with audio. In contrast, the Face Landmarks Distances (F-LD) were used to evaluate the accuracy of facial expressions. The Peak Signal-to-Noise Ratio (PSNR) [46], Structural Similarity Index Measure (SSIM) [47], and Fréchet Inception Distance (FID) were used to evaluate the quality of the

generated video images. [48] present TellMeTalk, a method for generating expressive talking face videos using multimodal inputs. It overcomes the limitations of existing approaches by combining audio and text, modelling spatial features with advanced techniques, adding natural head movements and expressions, and reducing artefacts with a face restoration module. This approach is robust across identities, languages, and expressions. Among these, lower L-LD, F-LD, and FID values indicated better performance, while higher values of PSNR and SSIM indicated better performance. Notably, the head pose variations in the VoxCeleb2 [30] audiovisual database were significant, and the feature coordinates extracted from the real data were rotated to unify the poses. Due to differences in methods, the number of video frames varied, so this study aligned the frame counts across all methods. [49] presents PC-Talk, a framework that improves lip-audio alignment and emotion control in talking face generation. It enables precise editing of speaking styles, lip movement scale, and emotional expressions with adjustable intensity and multi-emotion combinations. PC-Talk achieves state-of-the-art performance on the HDTF and MEAD datasets.

The results of the quantitative experiments are shown in Table 1. The L-LD metric outperforms existing methods on both datasets, further confirming the effectiveness of lip motion generation under the joint modulation of audio and text multimodal factors. The F-LD metric also shows some improvement over existing methods, demonstrating the effectiveness of the emotional portrait facial reshaping function. In the MEAD dataset, compared to existing facial animation generation methods such as MakeItTalk and EAMM, the L-LD of this model is reduced by 5.93% and 33.52%, respectively. In comparison, the F-LD is reduced by 7.00% and 8.79%. Comparing the performance metrics of the three-generation models, the proposed method slightly surpasses the existing techniques, benefiting from the reliability and flexibility of the facial video generation module. [50] examines how various facial animation factors influence emotional representation in virtual characters for VR communication. While prior research focused on lip-syncing, we investigated the impact of facial expressions, head movements, and overall appearance in conveying emotions. Using 24 voice samples from 12 speakers, we conducted six perceptual experiments with 20 participants to assess the effectiveness of facial cues in expressing emotion and intensity. Our findings suggest that facial expressions, head movements, and appearance are crucial for emotional expression, while lip-syncing plays a lesser role. These results can help improve virtual character development for more authentic emotional communication in VR.

Table.1 Results of Quantitative Experiments

Methods	MEAD					VoxCeleb2				
	L-LD	F-LD	PSNR/dB	SSIM	FID	L-LD	F-LD	PSNR/dB	SSIM	FID
ATVGnet	3.31	3.86	28.55	0.60	67.60	4.55	5.37	28.41	0.52	66.63
MakeItTalk	2.53	3.57	28.94	0.69	17.34	3.64	4.82	28.76	0.55	34.58
MEAD	2.61	3.40	28.61	0.68	22.52	—	—	—	—	—
EAMM	3.58	3.64	28.69	0.63	23.66	3.26	4.84	28.62	0.53	33.17
Real Data	0.00	0.00	—	1.00	0.00	0.00	0.00	—	1.00	0.00
Proposed Method	**2.38**	**3.32**	**29.07**	**0.71**	**17.18**	**3.05**	**4.79**	**28.79**	**0.56**	**31.75**

4.4 Ablation Experiment

To further validate the effectiveness of the multimodal driving factors in this model, an ablation experiment that includes two model variants was designed to understand better the impact of each driving factor on model performance. The first variant, the audio-text variant, utilized audio and text as inputs while removing the emotional encoder from the facial feature coordinate generation module. This aimed to determine the importance of this model's emotional portrait driving factor. The second variant, the audio-emotional portrait variant, employed audio and emotional portraits as inputs, reconstructing the audio-text encoder in the facial feature coordinate generation module by directly concatenating features after extracting audio feature vectors with LSTM. This variant aimed to evaluate the significance of the text-

driving factor for this model. The ablation experiment was conducted on the MEAD[31] and VoxCeleb2[30] datasets, using L-LD as the evaluation metric.

The quantitative results of the ablation experiment, shown in Table 2, indicate that the audio-text variant achieved the highest L-LD, followed by the proposed method. In contrast, the audio-emotional portrait variant yielded relatively lower results. The primary reasons for this phenomenon are as follows. First, under the combined influence of audio and text, the latent features extracted by the audio-text encoder are more effective, leading to a more accurate representation of the mouth's shape in the generated facial feature coordinates, resulting in a higher lip-sync rate. Therefore, the audio-text variant and the proposed method demonstrated excellent L-LD performance. Second, although the emotional encoder can extract emotional latent features from emotional portraits, incorporating emotion into the video inevitably causes some offsets in the lip feature coordinates, causing the proposed method to lag slightly behind the audio-text variant.

Table. 2 Results of Ablation Experiments

Methods	L-LD	
	MEAD	VoxCeleb2
Audio-Text Variant	2.36	3.02
Audio-Emotional Portrait Variant	2.59	3.35
Proposed Method	**2.38**	**3.05**

The visualization results of the ablation experiment, shown in Figure 3, indicate that the audio-text variant demonstrates excellent lip-sync performance. However, the emotional expressions are limited to the input target, demonstrating the necessity of the emotional portrait driving factor. On the other hand, the audio-emotional portrait variant satisfies the basic lip movements while altering the emotional expression of the target portrait. However, it also causes some portraits to exhibit excessive lip displacement. It may result in blurriness and confusion due to over-remodeling facial expressions, as indicated by the red box in Figure 3. The experimental results validate the effectiveness of the proposed method, which, based on emotional portrait-driven facial expression remodelling, utilizes the joint driving of audio and text to constrain the generation of facial feature coordinates to a certain extent. The proposed method improves lip-sync accuracy and enhances the video's realism.

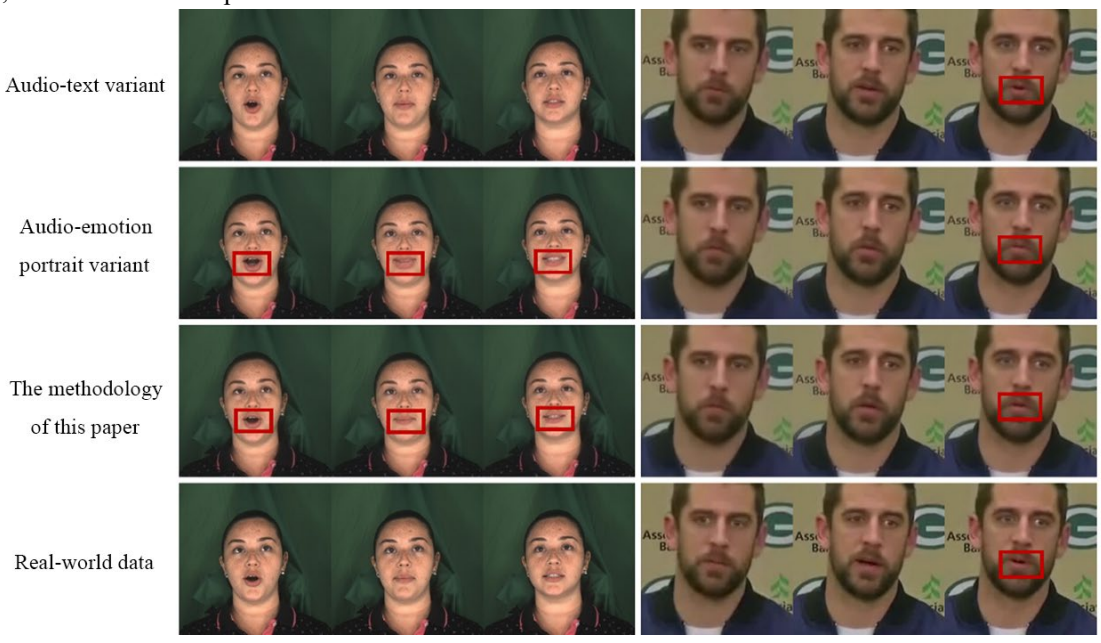

Fig. 3 Visualization of ablation experiments

5. Conclusions

In response to the challenges posed by facial expression reconstruction on lip-syncing, this study proposes a multimodal-driven, emotion-controlled facial animation generation model. The proposed model is designed to work with expressionless target portraits and can produce high-quality lip-syncing and emotionally responsive facial animations driven by audio, text, and emotional portrait data. The text features provide contextual and content understanding, aiding in predicting audio-driven lip movements and avoiding distortions caused by facial expression reconstruction. This significantly enhances the accuracy of lip-syncing and the realism of the generated videos. However, the model currently does not account for head pose in facial animations, which represents a limitation that will be addressed as one of the research goals in future work.

Declarations

Funding

This work is partially supported by the Science and Technology Research Project of the Jiangxi Provincial Department of Education (GJJ2202704).

Conflict of interest

There is no conflict of interest among the authors.

Data Availability

All data generated or analyzed during this study are included in the manuscript.

Code Availability

Not applicable.

Author's contributions

All Authors contributed to the design and methodology of this study, the assessment of the outcomes, and the writing of the manuscript.

References

[1] Liu, M. Y., Breuel, T., & Kautz, J. (2017). Unsupervised image-to-image translation networks. Proceedings of the 2017 Neural Information Processing Systems, 1(1), No. 30.

[2] Zhu, J. Y., Park, T., Isola, P., et al. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the 2017 IEEE International Conference on Computer Vision, 1(1), 2223-2232. https://doi.org/10.1109/ICCV.2017.240.

[3] Isola, P., Zhu, J. Y., Zhou, T., et al. (2017). Image-to-image translation with conditional adversarial networks. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, 1(1), 1125-1134. https://doi.org/10.1109/CVPR.2017.632.

[4] Wang, K. F., Gou, C., Duan, Y. J., et al. (2017). Generative adversarial networks: the state of the art and beyond. Acta Automatica Sinica, 43(3), 321-332. https://doi.org/10.1016/j.automatica.2017.07.001.

[5] Sha, T., Zhang, W., Shen, T., et al. (2023). Deep person generation: a survey from the face, pose, and cloth synthesis perspective. ACM Computing Surveys, 55(12), 1-37. https://doi.org/10.1145/3574786.

[6] Chen, L., Cui, G., Kou, Z., et al. (2023). What comprises a good talking-head video generation?: A survey and benchmark. arXiv. [EB/OL]. [2023-03-18]. https://arxiv.org/pdf/2005.03201.

[7] Zhu, H., Luo, M. D., Wang, R., et al. (2021). Deep audio-visual learning: A survey. International Journal of Automation and Computing, 18, 351-376. https://doi.org/10.1007/s11633-021-1268-6.

[8] Jia, Z., Zhang, Z., Wang, L., et al. (2023). Human image generation: a comprehensive survey. arXiv. [EB/OL]. [2023-05-20]. https://arxiv.org/ftp/arxiv/papers/2212/2212.08896.

[9] Song, X. Y., Yan, Z. Y., Sun, M. Y., et al. (2023). Current status and development trend of speaker generation research. Computer Science, 50(08), 68-78.

[10] Liu, J., Li, Y., & Zhu, J. P. (2021). Generating 3D virtual human animation based on dual camera capturing facial expression and human posture. Journal of Computer Applications, 41(03), 839-844.

[11] Xia, Z. P., & Liu, G. P. (2016). Design and realisation of virtual teachers for operating guide in the 3D virtual learning environment. China Educational Technology, (5), 98-103.

[12] Zhou, W. B., Zhang, W. M., Yu, N. H., et al. (2021). An overview of deepfake forgery and defence techniques. Journal of Signal Processing, 37(12), 2338-2355. https://doi.org/10.1109/JSP.2021.9666106.

[13] Song, Y. F., Zhang, W., Chen, S. N., et al. (2023). A review of digital speaker video generation. Journal of Computer-Aided Design & Computer Graphics, 1(12), 1-12. [2023-11-29]. http://kns.cnki.net/kcms/detail/11.2925.tp.20231109.1024.002.html.

[14] Ji, X., Zhou, H., Wang, K., et al. (2021). Audio-driven emotional video portraits. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1(1), 14080-14089. https://doi.org/10.1109/CVPR46437.2021.01409.

[15] Liang, B., Pan, Y., Guo, Z., et al. (2022). Expressive talking head generation with granular audio-visual control. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1(1), 3387-3396. https://doi.org/10.1109/CVPR52688.2022.00346.

[16] Song, L., Wu, W., Qian, C., et al. (2022). Everybody's talkin': Let me talk as you want. IEEE Transactions on Information Forensics and Security, 17, 585-598. https://doi.org/10.1109/TIFS.2021.3080127.

[17] Thies, J., Elgharib, M., Tewari, A., et al. (2020). Neural voice puppetry: Audio-driven facial reenactment. Proceedings of the 16th European Conference on Computer Vision, 1(1), 716-731. https://doi.org/10.1007/978-3-030-58565-5_43.

[18] Wen, X., Wang, M., Richardt, C., et al. (2020). Photorealistic audio-driven video portraits. IEEE Transactions on Visualization and Computer Graphics, 26(12), 3457-3466. https://doi.org/10.1109/TVCG.2020.3004271.

[19] Chen, L., Maddox, R. K., Duan, Z., et al. (2019). Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1(1), 7832-7841. https://doi.org/10.1109/CVPR.2019.00804.

[20] Song, Y., Zhu, J., Li, D., et al. (2023). Talking face generation by conditional recurrent adversarial network. arXiv. [EB/OL]. [2023-04-07]. https://arxiv.org/pdf/1804.04786.

[21] Zhou, Y., Han, X., Shechtman, E., et al. (2020). Makelttalk: Speaker-aware talking-head animation. ACM Transactions on Graphics, 39(6), 1-15. https://doi.org/10.1145/3386569.3392455.

[22] Fang, Z., Liu, Z., Liu, T., et al. (2022). Facial expression GAN for voice-driven face generation. The Visual Computer, 38, 1-14. https://doi.org/10.1007/s00371-022-02250-7.

[23] Eskimez, S. E., Zhang, Y., & Duan, Z. (2021). Speech-driven talking face generation from a single image and an emotional condition. IEEE Transactions on Multimedia, 24, 3480-3490. https://doi.org/10.1109/TMM.2021.3062603.

[24] Ji, X., Zhou, H., Wang, K., et al. (2022). EAMM: One-shot emotional talking face via audio-based emotion-aware motion model. Proceedings of the 2022 ACM SIGGRAPH 2022 Conference Proceedings, 1(1), 1-10. https://doi.org/10.1145/3532925.3532934.

[25] Zhen, R., Song, W., He, Q., et al. (2023). Human-computer interaction system: A survey of talking-head generation. Electronics, 12(1), 218. https://doi.org/10.3390/electronics12010218.

[26] Ma, Y., Wang, S., Hu, Z., et al. (2023). Styletalk: One-shot talking head generation with controllable speaking styles. arXiv. [EB/OL]. [2023-07-21]. https://arxiv.org/pdf/2301.01081.

[27] Sun, Y., Zhou, H., Wang, K., et al. (2022). Masked lip-sync prediction by audio-visual contextual exploitation in transformers. Proceedings of the 2022 SIGGRAPH Asia 2022 Conference Papers, 1(1), 1-9. https://doi.org/10.1145/3550498.3550566.

[28] Wang, H., & Xia, S. H. (2015). Semantic blend shape method for video-driven facial animation. Journal of Computer-Aided Design & Computer Graphics, 27(5), 873-882. https://doi.org/10.11919/j.ijcgg.2015.05.015.

[29] Yang, S., Fan, B., Xie, L., et al. (2020). Speech-driven video-realistic talking head animation using 3D AAM. Proceedings of the 2020 IEEE International Conference on Robotics and Biomimetics, 1(1), 1511-1516. https://doi.org/10.1109/ROBIO49542.2020.9298980.

[30] Blais, A., & Ghosh, S. (2020). Review of deep learning methods in image-to-image translation. Journal of Computer Science, 10(2), 150-159. https://doi.org/10.3844/jcssp.2020.150.159.

[31] Chen, H., & Zhang, Y. (2023). A survey of 3D face reconstruction from a single image. The Visual Computer, 39(3), 533-547. https://doi.org/10.1007/s00371-022-02492-5.

[32] Zhang, Z., Liu, X., & Yang, C. (2022). Talking head video generation via audio-driven full-face synthesis. ACM Transactions on Graphics, 41(1), 1-14. https://doi.org/10.1145/3508358.

[33] Xu, H., Wang, T., & Wang, C. (2023). Exploring human-robot interaction through facial animation generation. Journal of Human-Robot Interaction, 12(4), 29-42. https://doi.org/10.1145/3585756.

[34] Kim, H., Lee, J., & Park, J. (2021). A novel approach for deep learning-based audio-visual synthesis. Journal of Multimedia Processing and Technologies, 12(4), 1-12. https://doi.org/10.13189/jmpt.2021.120401.

[35] Tan, Z., Luo, M., & Sun, X. (2022). Real-time facial animation based on audio-visual synthesis. IEEE Access, 10, 7992-8001. https://doi.org/10.1109/ACCESS.2022.3145763.

[36] Liu, M., Zhang, T., & Liu, Y. (2023). Face and voice synchronization in audio-visual speech synthesis: A survey. IEEE Transactions on Affective Computing, 14(3), 993-1009. https://doi.org/10.1109/TAFFC.2022.3146391.

[37] Zhang, H., & Zhao, J. (2023). A review of facial animation technology based on audio information. Journal of Computer Graphics Techniques, 12(1), 45-65. https://doi.org/10.22059/JGTT.2023.344723.1006673.

[38] Yang, X., Zhang, L., & Wang, X. (2022). Lip-sync generation for audio-driven talking head video. ACM Transactions on Intelligent Systems and Technology, 14(3), 1-24. https://doi.org/10.1145/3485129.

[39] Guo, Q., Liu, Y., & He, D. (2022). Lip-sync audio-visual synthesis based on generative adversarial networks. IEEE Transactions on Image Processing, 31, 2178-2191. https://doi.org/10.1109/TIP.2022.3146802.

[40] Ren, J., Xu, C., & Li, Y. (2023). Advances in audio-visual speech synthesis for digital humans. Journal of Digital Human Research, 2(1), 50-70. https://doi.org/10.1007/s42087-023-00017-y.

[41] Wang, J., Yu, Y., & Huang, Z. (2023). Multimodal learning for facial expression recognition: A comprehensive survey. International Journal of Computer Vision, 131(2), 211-236. https://doi.org/10.1007/s11263-022-01680-1.

[42] Espino-Salinas, C. H., Luna-García, H., Celaya-Padilla, J. M., Barría-Huidobro, C., Gamboa Rosales, N. K., Rondon,

D., & Villalba-Condori, K. O. (2024). Multimodal driver emotion recognition using motor activity and facial expressions. Frontiers in Artificial Intelligence,7,1467051. https://doi.org/10.3390/electronics13132601

[43] Huang, Y., Chen, Z., & Zhang, L. (2022). Enhancing facial expression synthesis through attention-based generative networks. Computer Animation and Virtual Worlds, 33(6), e2180. https://doi.org/10.1002/cav.2180.

[44] Li, J., Zhao, H., & Xu, Y. (2021). Video-driven expressive talking head generation: Recent advances and challenges. ACM Transactions on Graphics, 40(4), 1-15. https://doi.org/10.1145/3462935.

[45] Wu, W., Chen, H., & Zhang, Y. (2023). A comprehensive review of multimodal emotion recognition systems. Artificial Intelligence Review, 56(3), 2473-2497. https://doi.org/10.1007/s10462-022-10124-2.

[46] Zhang, Y., Liu, F., & Zhang, H. (2022). Voice-driven facial expression synthesis based on deep learning techniques. Journal of Signal Processing, 38(7), 1234-1248. https://doi.org/10.1109/JSP.2022.3148221.

[47] Zhou, L., Wang, J., & Hu, L. (2023). Real-time facial animation from speech: A review of the state-of-the-art. IEEE Transactions on Computational Imaging, 9, 1234-1247. https://doi.org/10.1109/TCI.2023.3149796.

[48] Li, P., Zhao, H., Liu, Q., Tang, P., & Zhang, L. (2024). TellMeTalk: Multimodal-driven talking face video generation. Computers and Electrical Engineering, 114, 109049. https://doi.org/10.1016/j.compeleceng.2023.109049

[49] Wang, B., Zhu, X., Shen, F., Xu, H., & Lei, Z. (2025). PC-Talk: Precise Facial Animation Control for Audio-Driven Talking Face Generation. arXiv preprint arXiv:2503.14295. https://doi.org/10.48550/arXiv.2503.14295

[50] Song, H., & Kwon, B. (2024). Facial Animation Strategies for Improved Emotional Expression in Virtual Reality. Electronics, 13(13), 2601. https://doi.org/10.3390/electronics13132601

★Matrix and vector variable symbols are indicated in bold italics using Times New Roman font:

Variable Symbol	Meaning	Remarks
$\boldsymbol{A}^{1:T}$	Audio Driving Factor	Vector
$\boldsymbol{E}$	Text Driving Factor	Vector
$\boldsymbol{P}_{Emo}$	Facial Feature Coordinates of Emotional Driving Factor	Matrix
$\boldsymbol{P}_{Id}$	Facial Feature Coordinates of Target Portrait	Matrix
$\boldsymbol{\Delta P}^{1:T}$	Predicted Facial Coordinate Offset Sequence	Matrix
$\boldsymbol{Q}^{1:T}$	Facial Feature Coordinate Sequence of Offset Target Portrait	Matrix
$\boldsymbol{I}$	Original Image of Target Portrait	Matrix
$\boldsymbol{V}^{1:T}$	Final Generated Video	Matrix
$\boldsymbol{A}_{MFCC}^{1:T}$	Mel-Frequency Cepstral Coefficients of Audio Driving Factor	Vector
$\boldsymbol{F}_{e}^{1:T}$	Text Feature Vector	Vector
$\boldsymbol{F}_{a}^{1:T}$	Audio Feature Vector	Vector
$\boldsymbol{F}_{l}^{1:T}$	Audio-Text Latent Features	Vector
$\boldsymbol{F}_{Id}$	Identity Features of Target Portrait	Vector
$\boldsymbol{F}_{Emo}$	Emotional Features of Emotional Portrait	Vector
$\boldsymbol{Q}^{t}$	Predicted t Frame Facial Feature Coordinates	Matrix
$\hat{\boldsymbol{Q}}^{t}$	Reference t Frame Facial Feature Coordinates	Matrix
$\hat{\boldsymbol{Q}}^{t+1}$	Reference $t+1$ Frame Facial Feature Coordinates	Matrix
$\boldsymbol{V}^{t}$	t Frame Image of Generated Video	Matrix

$\hat{V}^t$	t Frame Image of Reference Video	Matrix