







# EAPC: Emotion and Audio Prior Control Framework for the Emotional and Temporal Talking Face Generation

Xuan-Nam Cao<sup>1,2</sup><sup>a</sup>, Quoc-Huy Trinh<sup>1,2</sup><sup>b</sup>, Quoc-Anh Do-Nguyen<sup>1,2</sup><sup>c</sup>, Van-Son Ho<sup>1,2</sup><sup>d</sup>,  
Hoai-Thuong Dang<sup>1,2</sup><sup>e</sup> and Minh-Triet Tran<sup>1,2</sup><sup>f</sup>

<sup>1</sup>Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam

<sup>2</sup>Vietnam National University, Ho Chi Minh City, Vietnam  
fi

**Keywords:** Landmark Generation, Talking Head, Dual-LSTM, Acoustic Features.

**Abstract:** Generating realistic talking faces from audio input is a challenging task with broad applications in fields such as film production, gaming, and virtual reality. Previous approaches, employing a two-stage process of converting audio to landmarks and then landmarks to a face, have shown promise in creating vivid videos. However, they still face challenges in maintaining consistency due to misconnections between information from the previous audio frame and the current audio frame, leading to the generation of unnatural landmarks. To address this issue, we propose EAPC, a framework that incorporates features from previous audio frames with the current audio feature and the current facial landmark. Additionally, we introduce the Dual-LSTM module to enhance emotion control. By doing so, our framework improves the temporal aspects and emotional information of the audio input, allowing our model to capture speech dynamics and produce more coherent animations. Extensive experiments demonstrate that our method can generate consistent landmarks, resulting in more realistic and synchronized faces, leading to the achievement of our competitive results with state-of-the-art methods. The implementation of our method will be made publicly available upon publication.


## 1 INTRODUCTION


Generating a talking face from audio involves creating a realistic face based on the audio input and a reference face to be generated. This technique holds significant value with widespread applications in film, gaming, virtual reality, education, and communication. It contributes to enhanced immersion, storytelling, and character development by precisely synchronizing facial expressions with audio cues.


With the advancement of Deep Learning, previous works have demonstrated notable performance in generating realistic faces. Recently MEAD (Wang et al., 2020), and CREMA-D (Cao et al., 2014) have been introduced a high-quality audio-visual dataset and a pipeline for high-quality talking face genera-


tion. Since then, several methods have been proposed to address this challenge, and these approaches can be divided into two main categories. Firstly, several approaches such as Speech2Vid (Chung et al., 2017), Conditional Recurrent Adversarial Network (Song et al., 2019), the method proposed by Mittal et al. (Mittal and Wang, 2020), and Sinha et al. approach try to animate one or few frames of cropped. The second group of approaches are video-based editing methods such as EVP (Ji et al., 2021), SSP-NeRF (Liu et al., 2022), and EAMM (Ji et al., 2022) strive to directly edit target video clips faces of. All of these methods showcase their capability to explicitly control emotions in the upper face, yielding impressive results in facial emotion generation, and enhancing the quality of image generation.


Although previous methods achieve some prominent results, they are focusing on the improvement of the facial quality, and the integration of emotion, still encounter certain challenges: (1) the inconsistency of the frames during the long frames, which affects to the temporal characteristic of the face animation, (2) the lack of focusing on the informative parts of the face


<sup>a</sup> <https://orcid.org/0000-0002-3614-7982>

<sup>b</sup> <https://orcid.org/0000-0002-7205-3211>

<sup>c</sup> <https://orcid.org/0009-0006-7664-4613>

<sup>d</sup> <https://orcid.org/0000-0002-8389-2176>

<sup>e</sup> <https://orcid.org/0009-0009-2127-1364>

<sup>f</sup> <https://orcid.org/0000-0003-3046-3041>

in previous methods can lead to the missing of facial texture. These issues can collectively contribute to the unnatural appearance of animated faces in applications.

To overcome previous challenges, we propose a framework named **Emotion and Audio Prior Control (EAPC)**, an image-based method, for the emotional and temporal Talking Face Generation. This framework can selectively focus on relevant facial landmarks, considering contextual information from both the current and previous frames. As a result, this method can keep the consistency of the animated faces and the temporal characteristic of the sequence, thereby aiding in the generation of a natural animated face. Furthermore, we also propose our Dual-LSTM module, which enables our approach’s ability to select the informative landmark information correlated with the emotion target, enhancing our framework’s control over landmark generation based on the target emotion.

In summary, there are two main contributions:

- We propose Emotion and Audio Prior Control, the framework that control the Talking face generation by using the emotion and prior audio frame information for improving the talking face generation.
- We introduce the Dual-LSTM module, which facilitates the fusion of landmark and audio features. This module also possesses the capability to selectively choose informative landmarks based on the control of the target emotion.

The remainder of this paper is organized as follows. Section 2 provides an overview of related work in the field of realistic facial animation generation. Section 3 describes the proposed methodology, including the details of the EAPC framework, and Dual-LSTM module. Section 4 presents the experimental setup, datasets, and evaluation metrics used in our study. The experimental results and analysis are presented in Section 5. Finally, Section 6 concludes the paper and discusses potential future research directions.

## 2 RELATED WORK

### 2.1 Talking Face Generation

Talking Face Generation is the technique that aims to generate animated faces via the audio of talk. In recent years, there has been a growing interest in generating photo-realistic talking faces using deep learn-

ing techniques. There are two main categories of approaches: image-based methods and video editing-based methods.

#### 2.1.1 Image-Based Method

This method focuses on generating an animated face based on reference facial images. One of the pioneering approaches in this category is Speech2Vid (Chung et al., 2017), which utilizes a combination of audio encoder and Identity Reference encoder to generate lip-sync videos in an image-to-image translation manner (Isola et al., 2017). Subsequently, both (Zhou et al., 2019) and (Song et al., 2018) employ adversarial learning with disentangled audio-visual representation to enhance the model’s learning through joint embeddings. Following these efforts, (Chen et al., 2019) introduces ATVGnet, a novel cascade network structure that integrates emotion into video generation. They also explore the use of pixel-wise loss with an attention mechanism to enhance the temporal aspects of the generated video. This marks the first instance where the temporal characteristics of the generated video are considered, paving the way for subsequent research.

Since then, numerous methods have been proposed to address the challenge of incorporating emotion. A notable method is presented by (Vougioukas et al., 2019), who employ three separate discriminators to enhance synthesis details, synchronization, and realistic expressions. More recently, (Wang et al., 2020) introduced the MEAD dataset along with a pipeline for emotional talking face generation. However, this method demonstrates a lack of temporal coherence and tends to produce unnatural emotions in the generated faces. This limitation has prompted later methods to address the temporal aspect, such as the pipeline for facial geometry-aware one-shot emotional talking face generation from audio with independent emotion control by (Sinha et al., 2022). Another method by (Eskimez et al., 2021) introduces an architecture with an emotion discriminative loss that classifies rendered visual emotions, fusing audio and reference images. However, this fusion architecture is conditioned by the emotion category.

While these methods show improvement in emotional talking face generation, they encounter challenges in preserving temporal characteristics when dealing with audio and a single image to create a sequential output.

#### 2.1.2 Video-Based Editing Methods

A video-based editing method involves using portraits that encompass not only the facial areas but also the

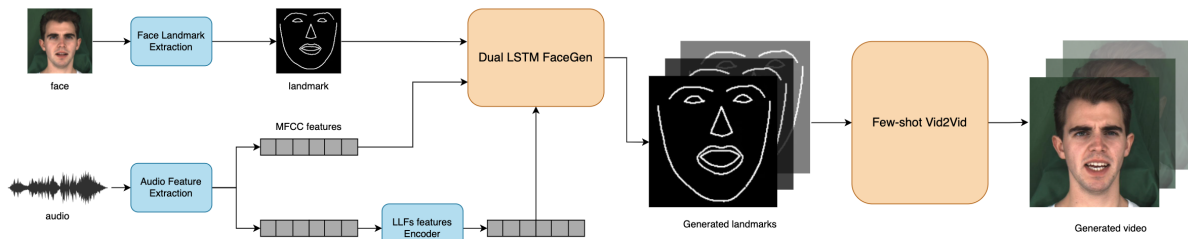


Figure 1: General visualization of EAPC framework.

neck and shoulder parts of a person, along with the background. These methods face challenges in generating realistic faces due to the intricate relationship between the face and other body parts. One noteworthy method in this category is presented by (Suwajanakorn et al., 2017), who synthesizes photorealistic talking videos of Obama by training an audio-to-landmark model using recurrent neural networks (RNN). Additionally, both (Song et al., 2018) and (Thies et al., 2020) regress facial expression parameters of 3DMM models and inpaint the mouth.

A more recent advancement in this field is the Emotional Video Portrait (EVP) method introduced by (Ji et al., 2021). EVP is the first to achieve emotional control in talking face generation, including the dynamic movement of the generated face. The evolution of these methods demonstrates promising potential for the utilization of emotional control in the generation of talking faces.

In general, previous works have demonstrated high-quality results, they still grapple with challenges related to the temporal aspects of the video and the incorporation of emotional information. In our work, our method aims to address these issues, ultimately enhancing the naturalness of the animated face. This endeavor opens up a promising avenue for advancements in this field.

## 2.2 Attention Mechanism

The Attention Mechanism, a renowned concept introduced by (Vaswani et al., 2017), has found application in Transformer models for various tasks related to sequence, sentence, and more recently, image generation. The fundamental idea behind this method revolves around considering different positions within a single sequence to compute a comprehensive representation of that sequence. By incorporating the attention mechanism, a sequence can maintain long-range dependencies, addressing a challenge faced by previous methods such as RNN and LSTM. Additionally, attention facilitates the identification and emphasis of correlations within the sequence, enabling the model to focus on selective features.

Recent works, exemplified by (Wang et al., 2022), (Wang et al., 2023b), and (Wang et al., 2023a), have integrated the attention module as part of the Transformer architecture for talking face generation. Their results showcase the effectiveness of attention in efficiently capturing information across long-range sequences and emphasizing crucial aspects of the face, ultimately enhancing the smoothness of the output video. Inspired by these ideas, our Dual-LSTM module incorporates the attention mechanism for efficient feature selection, building upon the success demonstrated by attention in similar applications.

## 3 EAPC FRAMEWORK

In this section, we present the EAPC framework. Our method is divided into two stages, as illustrated in Figure 1. In the first stage, we introduce the refined audio-to-landmark generation concept, which is based on the fusion of the reference face’s landmarks with the current and previous audio frames. Additionally, we propose the Dual-LSTM module for the audio and landmark fusion. Moving to the second stage, we leverage Few-shot Vid2Vid generation (Wang et al., 2019) for landmark-to-video generation.

### 3.1 Overview

As illustrated in Figure 1, the input to the EAPC framework consists of two pieces of information: the reference face and the audio sequence. The audio feature is initially extracted by the Audio extraction modules, capturing both Mel Frequency Cepstral Coefficients (MFCC) features and Low-Level Features (LLFs). Subsequently, landmarks are extracted from facial images using the Mediapipe framework (Lugaresi et al., 2019). Following this, emotional information is derived from the audio through our designed LLFs modules. Our proposed Dual-LSTM then integrates the landmark and audio features to generate a sequence of landmarks. To further modulate the emotion of the generated landmark during speech,

our Dual-LSTM incorporates the emotion audio feature with the landmark using the local attention block. In the second stage, we fine-tune the Vid2Vid model with the landmark output from the Dual-LSTM module, facilitating the generation of a sequence of talking face frames. We emphasize the significance of the Audio-to-landmark stage, as it directly influences the subsequent stage of generating the sequence of face images.

### 3.2 Audio Features Extraction

In order to capture the content and emotion of the audio input, we employ a process of audio feature extraction. One of the primary features is Mel Frequency Cepstral Coefficients (MFCC), which provides a representation of the spectral characteristics of the audio signal. Each frame of audio is transformed into a matrix of size  $28 \times 12$ . In addition to MFCC, we also extract a set of low-level features (LLFs) that capturing the emotional variations present in the audio signal. These LLFs include:

1. **Root Mean Square Error (RMSE):** RMSE measures the average magnitude of the audio signal. It reflects the overall energy and loudness of the audio.
2. **Chroma:** Chroma represents the distribution of musical pitch classes in the audio. It captures tonal information and can be useful for detecting emotional cues related to musical harmony.
3. **Spectral Centroid:** Spectral Centroid represents the center of mass of the power spectrum of the audio signal. It provides information about the brightness or timbre of the sound and can indicate emotional variations related to the spectral characteristics.
4. **Spectral Bandwidth:** Spectral Bandwidth measures the range of frequencies covered by the power spectrum. It provides information about the width or spread of the spectral content and can reflect emotional variations related to the richness or thinness of the sound.
5. **Spectral Rolloff:** Spectral Rolloff represents the frequency below which a specified percentage of the total spectral energy resides. It can indicate emotional cues related to the high-frequency content or brightness of the sound.
6. **Zero Crossing Rate (ZCR):** ZCR measures the rate at which the audio signal changes sign (from positive to negative or vice versa). It provides information about the temporal characteristics and can be related to emotional variations in the dynamics of the sound.

7. **Spectral Flatness:** Spectral Flatness measures the ratio between the geometric mean and the arithmetic mean of the power spectrum. It can indicate emotional cues related to the noisiness or tonality of the sound.

### 3.3 LLFs Encoder

To explicitly extract emotion from audio features, we utilize a combination of three LSTM layers, referred to as the LLFs encoder. The input to this encoder is the low-level features (LLFs), denoted as  $X_{llfs}$ . These features, forming a 25-dimensional vector, are then straightforwardly passed through the LLFs Encoder network to generate the emotional information output ( $X_{emotion}$ ) with eight elements representing the probability distribution of each emotion category. Equation 1 illustrates the process of this encoder.

$$X_{emotion} = LLFs\_encoder(X_{llfs}) \quad (1)$$

By incorporating the  $X_{llfs}$  into the encoding process, we can capture and represent the various low-level emotional cues present in the input audio. These LLFs, which encompass characteristics such as spectral centroid, zero crossing rate, and spectral flatness, provide valuable information about the audio's tonality, noisiness, and temporal dynamics.

### 3.4 Dual-LSTM Module

As mentioned earlier, landmarks play a crucial role in generating realistic animated faces. Moreover, the temporal characteristics and emotional information are two vital aspects for creating lifelike talking faces. This is precisely why we introduced the Dual-LSTM, designed to address the challenges associated with incorporating both temporal dynamics and emotional cues in the audio-to-landmark generation process.

Our module consists of multiple stages, corresponding to the number of landmark frames. The complete design of this module is illustrated in Figure 2. The input to this module includes the MFCC feature ( $X_{MFCC}^i$ ) with a shape of  $N \times 25 \times 12$  and the emotion information with a shape of  $N \times 25 \times 8$ , where 25 represents the number of frames. The output of this module is the sequence of landmarks, denoted as  $out_{lm}$ , with a shape of  $N \times 25 \times 68 \times 2$ .

In each stage  $i$ , features ( $X_{MFCC}^i$ ) are initially extracted from each frame of the input audio using the representation outlined in Equation 2. These features are subsequently input into a 3-layer unidirectional audio LSTM module to capture temporal dependencies and extract meaningful audio information. The

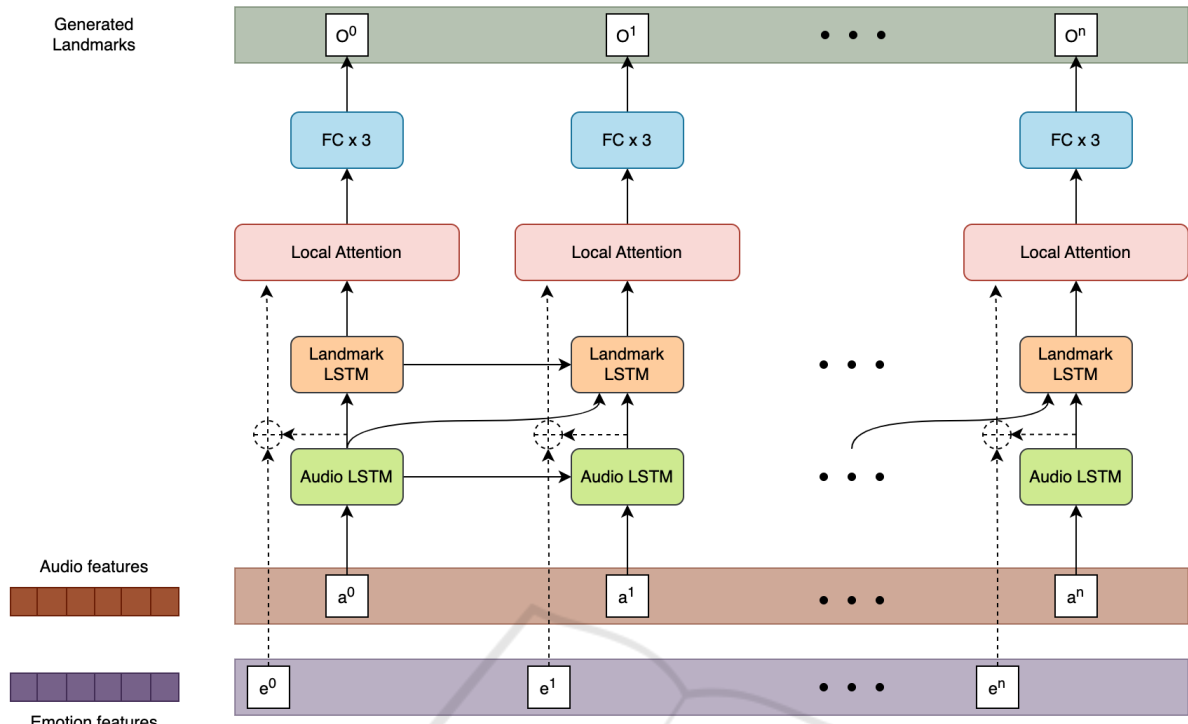


Figure 2: Design of the Dual-LSTM module.

resulting output contains encoded audio information for each time step in the input sequence.

$$X_{af}^i = LSTM_{audio}^i(X_{MFCC}^i) \quad (2)$$

The audio features are duplicated to create a controllable version ( $X_{afnew}^i$ ) influenced by the emotion feature through concatenation with emotional information, as depicted by Equation 3.

$$X_{afnew}^i = \text{concatenate}(X_{af}^i, X_{emotion}^i) \quad (3)$$

Subsequently, additional LSTM layers are employed to generate the landmark output. In this layer, a skip connection is established between the current frame and the previous frame, fostering the retention of temporal characteristics. Furthermore, this connection addresses the issue of insufficient long-range dependency inherent in LSTM modules. The process is represented by Equation 4, and Equation 5.

$$X_{lm}^i = LSTM_{lm}^i(X_{af}^i) + X_{lm}^{i-1}, (i > 0) \quad (4)$$

$$X_{lm}^i = LSTM_{lm}^i(X_{af}^i), (i = 0) \quad (5)$$

To regulate the output landmarks based on emotion and focus on selective features, we employ the local attention block. The inputs to this block are the audio feature in the duplicated version  $X_{afnew}^i$  and the generated landmark  $X_{lm}^i$ , leading to the production

of the controlled landmark  $output_{lm}^i$ . The subsequent equations elucidate the workings of the local attention mechanism:

$$att^i = \sigma(X_{afnew}^i \cdot X_{lm}^i) \cdot X_{lm}^i \quad (6)$$

$$output_{lm}^i = X_{afnew}^i + att^i \quad (7)$$

Which  $\sigma$  denotes softmax function

From Equations 6 and 7, the attention output is computed through the dot product of audio features and landmark features. Subsequently, the result undergoes weighting via the softmax function before undergoing another dot product with landmark features. Additionally, to incorporate context control into the attention feature, a skip connection is established between audio and landmark features. This skip connection ensures the retention of audio information.

Following this process, the output features capture the relevance and alignment among audio features, low-level features, and generated landmarks. These output features play a crucial role in guiding the refinement process and adjusting the landmarks accordingly. This ensures synchronization with the audio and accurate representation of the desired facial expressions.

Finally, a series of three fully connected layers with sizes of 512, 256, and 136 (68x2) are designed to generate accurate landmarks consisting of 68 two-dimensional points.

$$out\ put_{lm}^i = W_{138}(W_{256}(W_{512} * (out\ put_{lm}^i))) \quad (8)$$

In Equation 8,  $W_k$  with  $k \in \{136, 256, 512\}$  denote the learnable weight of these layers, which are responsible for generating the appropriate facial landmarks in the facial animation process. These layers are crucial for generating the appropriate facial landmarks during the facial animation process. Each fully connected layer plays a role in extracting higher-level representations from the input, gradually reducing the dimensions. Eventually, all the landmarks from each stage  $i$  are reshaped and concatenated to form the sequence of output landmarks with a shape of  $N \times 25 \times 68 \times 2$ , as depicted in Equation 9.

$$out\ put_{lm} = \text{concatenate}(\phi(out\ put_{lm}^0), \dots, \phi(out\ put_{lm}^{25})) \quad (9)$$

$\phi$  denotes the reshape function.

In general, by harnessing the connection between audio features and facial landmarks, the model seeks to enhance the temporal precision of landmark predictions and improve synchronization between audio and visual cues. Moreover, the incorporation of Dual-LSTM enables emotion to influence the landmarks, aiding in the expression of emotions in crucial facial regions. Consequently, this enhancement contributes to the generation of smooth and natural animated sequences of talking face landmarks.

### 3.5 Video Generator

In this research, we employ the few-shot Vid2Vid model (Wang et al., 2019) as a component in the process of generating realistic facial animations from audio input. The few-shot Vid2Vid model is a cutting-edge approach that enables the synthesis of visually coherent and synchronized facial animations by leveraging the power of deep learning techniques.

The few-shot Vid2Vid model operates by taking the extracted emotion features and the driving audio as input, and then performs a series of complex transformations to generate highly expressive and lifelike facial animations. It excels at capturing the intricate details of facial movements and effectively synchronizing them with the audio cues.

The input to this component includes the landmark sequence from the Dual-LSTM module ( $Out\ put_{lm}$ ) and the reference input face image  $X_{face}$ . The output of this module is the talking face video ( $X_{vid}$ ). We refer to the Vid2Vid module as  $f_{vid2vid}$ , and Equation 10 represents the integration of this module into our framework.

$$X_{vid} = f_{vid2vid}(X_{face}, Out\ put_{lm}) \quad (10)$$

By incorporating this component, we leverage its capability to generalize effectively to unseen or limited training data. It achieves this by utilizing few-shot learning techniques, enabling the model to learn from a small number of examples and generalize to new instances with high fidelity. This feature makes it particularly useful in scenarios where only a limited amount of training data is available.

Furthermore, the few-shot Vid2Vid model incorporates advanced techniques such as attention mechanisms and adversarial training, which further enhance the quality and realism of the generated facial animations. These techniques enable the model to focus on important regions of the face and effectively capture the dynamics of facial expressions.

### 3.6 Objective Function

**Audio to Landmark:** During the training process of the audio to landmark generation, we utilize the Mean Squared Error (MSE) loss and the landmark distance loss to compute the error metrics in our model.

The MSE loss (shown in Equation 11) is calculated as the average squared difference between the predicted landmarks  $Out\ put_{lm}$  and the ground truth landmarks  $GT_{lm}$ :

$$\mathcal{L}_{mse} = \frac{1}{N} \sum_{i=1}^N \|Out\ put_{lm}^i - GT_{lm}^i\|^2 \quad (11)$$

where  $N$  represents the total number of training samples, and  $i$  represents the index of landmark.

The landmark distance loss (shown in Equation 12) focuses on capturing the spatial relationships and relative distances between the predicted landmarks. It encourages the model to generate landmarks that are consistent with the structural characteristics of the face. The landmark distance loss is defined as:

$$\mathcal{L}_{lmd} = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^K \|Out\ put_{lm}^i - Out\ put_{lm}^j\| \quad (12)$$

Where  $K$  denotes the number of landmarks,  $i$  is the index of the landmark, and  $j$  denotes the index of landmark points.

The total loss of this stage is defined in the Equation 13

$$\mathcal{L}_{total} = \mathcal{L}_{lmd} + \mathcal{L}_{mse} \quad (13)$$

By minimizing these loss functions during training, our model learns to reduce the discrepancies between the predicted landmarks and the ground truth landmarks. These loss functions serve as important guiding principles, allowing the model to effectively

learn the intricate details and spatial relationships required for realistic facial animation generation.

**Landmark to Video:** For the landmark-to-video stage, we employ a combination of GAN loss as mentioned in Vid2Vid (Wang et al., 2019). By optimizing this objective function, our model can learn the weights necessary for generating a realistic face based on the input landmarks.

## 4 EXPERIMENTS

### 4.1 Datasets

In the experiment, to evaluate the efficiency, and the accuracy of EAPC, we use two datasets are MEAD (Wang et al., 2020) and CREMA-D (Cao et al., 2014), two public datasets for high quality talking face generation.

**MEAD Dataset:** (Wang et al., 2020) This dataset consists of a total of 281,400 high quality clips from 60 actors, including the emotional label, with a resolution of  $1920 \times 1080$ , and it is used for benchmark the talking head problems in several previous approaches.

**CREMA-D Dataset:** (Cao et al., 2014) This dataset comprises approximately 7,442 clips from 91 actors, with a resolution of  $1280 \times 720$ . This is one of the first datasets used for emotional talking face generation and emotion recognition.

### 4.2 Evaluation Metric

To quantitatively evaluate the performance of different methods, we conducted several metrics-based assessments. Firstly, facial landmarks were extracted from both the generated sequences and the ground truth sequences, with alignment performed to compensate for head motions. Landmark Distance (LD) and Landmark Velocity Difference (LVD) metrics (Chen et al., 2018), (Zhou et al., 2020) were employed to evaluate the accuracy of facial motions. LD represents the average Euclidean distance between the generated and recorded landmarks, while LVD quantifies the average velocity differences of landmark motions between the two sequences. Specifically, we focused on evaluating the lip movements and facial expressions separately by applying LD and LVD metrics to the mouth (M-LD, M-LVD) and face areas (F-LD, F-LVD).

Furthermore, to assess the image quality of the generated frames, we compared the Structural Similarity Index (SSIM) (Larkin, 2015), Peak Signal-to-Noise Ratio (PSNR), and Fréchet Inception Distance

(FID) (Heusel et al., 2018) scores. These metrics provided insights into the visual fidelity and realism of the synthesized images.

### 4.3 Implementation Detail

The configuration and parameters employed for the training of the EAPC framework on both the CREMA-D and MEAD datasets were specified as follows: In the case of the CREMA-D Dataset, the training phase involved 5,953 files, constituting 79,416 samples, while the testing phase comprised 1,489 files, encompassing 19,890 samples. Concerning the MEAD Dataset, the training dataset comprised 533 files, comprising 10,650 samples, and the testing dataset included 134 files, with 2,547 samples. All of the experiments are done on the video with 25 frames per second (FPS), and the Mediapipe framework (Lugaresi et al., 2019) is used to extract facial landmarks from the video. The hyperparameters applied to both datasets encompassed 500 epochs, a batch size of 32, and a learning rate set at  $1.0 \times 10^{-4}$ . The deliberate selection of these configurations and parameters was geared towards ensuring the effective training and evaluation of the models within the context of their respective datasets.

## 5 RESULT

### 5.1 Comparison Methods

We conduct a comparative analysis of the proposed EAPC framework with cutting-edge approaches in the domain of talking face generation. The evaluation is performed on the MEAD dataset and CREMA-D dataset, featuring benchmarks from leading methodologies such as Emotion-controllable generalized talking face generation by Sinha et al. (Sinha et al., 2022), Audio-driven emotional video portraits by Ji et al. (Ji et al., 2021), Mead: A large-scale audio-visual dataset for emotional talking-face generation by Kaisiyuan et al. (Wang et al., 2020), Realistic speech-driven facial animation with GANs by Vougioukas et al. (Vougioukas et al., 2019), and Speech-driven talking face generation from a single image and an emotion condition by Eskimez et al. (Eskimez et al., 2021).

### 5.2 Qualitative Result

Table 1 presents a comparison between our method and state-of-the-art approaches. It is evident that our method outperforms others in the MEAD dataset,

Table 1: Qualitative results landmark quality and texture quality for different methods on two datasets (MEAD and CREMA-D).

Dataset	Method	Landmark quality				Texture Quality		
		M-LD ↓	M-LVD ↓	F-LD ↓	F-LVD ↓	PSNR ↑	SSIM ↑	FID ↓
MEAD	MEAD (Wang et al., 2020)	2.52	2.28	3.16	2.01	28.61	0.68	22.52
	EVP (Ji et al., 2021)	2.45	1.78	3.01	1.56	29.53	0.71	<b>7.99</b>
	(Sinha et al., 2022)	2.18	<b>0.77</b>	<b>1.24</b>	<b>0.50</b>	30.06	0.77	35.41
	<b>EAPC</b>	<b>2.01</b>	1.25	1.85	1.24	<b>32.03</b>	<b>0.79</b>	22.67
CREMA-D	(Vougioukas et al., 2019)	2.90	<b>0.42</b>	2.80	<b>0.34</b>	23.57	0.70	71.12
	(Eskimez et al., 2021)	6.14	0.49	5.89	0.40	30.91	0.85	218.59
	(Sinha et al., 2022)	2.41	0.69	1.35	0.46	31.07	<b>0.90</b>	68.45
	<b>EAPC</b>	<b>1.45</b>	1.16	<b>1.33</b>	1.11	<b>33.49</b>	0.855	<b>17.31</b>

demonstrating higher results in the MLD, PSNR, and SSIM metrics. Similarly, on the CREMA-D dataset, our method surpasses competitors in the MLD, FLD, PSNR, and FID metrics.

The analysis of the benchmark results in Table 1 reveals the superior performance of our proposed method across various evaluation metrics. Specifically, on the MEAD dataset, our method excels in landmark distance measurements (MLD) for both the mouth and facial regions, showcasing enhanced accuracy in capturing landmark locations. Moreover, our approach attains a higher PSNR value, indicating superior preservation of image details and reduced noise distortion. The elevated SSIM score further emphasizes our method’s ability to accurately retain overall texture patterns and structures.

Turning our attention to the evaluation of the CREMA-D dataset, our method stands out in terms of the PSNR and FID compared to other approaches. The achievement of PSNR metrics shows the high quality of face animated output of our method, compared to the ground truth. The favorable FID score suggests a closer match between the generated textures and the real textures in the CREMA-D dataset, underscoring the effectiveness of our approach in producing textures that closely resemble the ground truth. These results show the realism of the generated face, thus making the result more consistent and natural.

### 5.3 Qualitative Visualization

**Landmark Generation Result:** To evaluate the temporal of the generated landmarks, we do the visualization in the consecutive animated face, with sample from the test dataset, present from left to right and up to down. The landmarks are drawn with the yellow lines for the better view. The visualization can be seen in the Figure 3.

Observations from Figure 3 indicate that, in some instances, the predicted landmarks display continuity, and natural head motion, reflecting the tempo-

ral coherence of the generated landmark sequence. This result demonstrates the effectiveness of our innovative approach in maintaining the temporal consistency of landmark sequence generation. The employed methodology proves to be efficient in addressing and resolving this issue.

Furthermore, we present the faces generated by our comprehensive framework alongside the ground truth to visually evaluate the effectiveness of our method in illustrating the impact of temporal landmark generation and the incorporation of emotional information on the animated face result. Figure 4 provides a visual representation of our comparison.

**Face Generation Result:** We do the visualization for the face generation, and make the visual comparison between our result and the groundtruth, which can be seen in the Figure 4.

The visualization from Figure 4 clearly illustrates that the combination of the temporal characteristics from the generated landmark and the incorporation of emotional information results in a generated face with similar emotions. The output is consistent, maintaining high quality comparable to the ground truth. This observation highlights the effectiveness of our EAPC framework in producing high-quality emotional talking face videos characterized by natural and accurate lip movements, head movements, and expressions. Consequently, our framework holds promise for various applications.

### 5.4 Ablation Study

To evaluate the effectiveness of EAPC, we perform 2 studies: (1) the utilization of audio information from the previous audio step  $i-1$  to compare the performance between the model variants that considered and ignored the audio information from the previous step, and (2) the inclusion of an attention mechanism for comparing the model variants with and without the attention mechanism which allows the model to selectively attend to different parts of the input se-





Figure 3: The visualization of the talking landmark sequence.

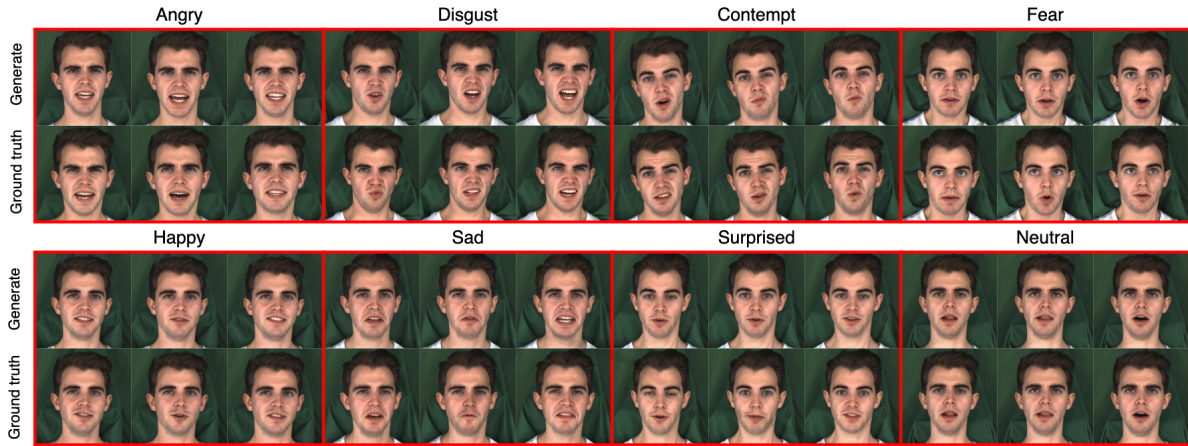


Figure 4: Comparison between our generated facial images and ground truth facial images.

quence, providing additional contextual information for the prediction process.

### 5.4.1 Prior Audio Frame Impact

In this study, we conducted experiments to assess the model’s performance with and without implementing skip connections to the prior frame (we evaluate without integrate emotional information with attention mechanism). The results, as presented in Table 2, offer a comparison between the methods that utilize the previous audio frame and those that do not.

Table 2: Comparison of methods with and without utilizing previous audio frame.

Method	M-LD ↓	M-LVD ↓	F-LD ↓	F-LVD ↓
w/o Audio i-1	2.18	1.38	1.99	1.35
w Audio i-1	<b>2.11</b>	<b>1.35</b>	<b>1.97</b>	<b>1.29</b>

From Table 2, it is evident that utilizing the previous audio frame (w Audio i-1) significantly enhances the measurements for M-LD, M-LVD, F-LD, and F-LVD, resulting in values of 2.11, 1.35, 1.97, and 1.29, respectively. These findings illustrate that the inclusion of the previous audio frame contributes to the temporal improvement in the landmark sequence, consequently leading to enhanced performance in these landmark evaluation metrics.

### 5.4.2 Impact of Attention in Emotion Control

In this study, we conducted experiments to assess the effectiveness of the attention mechanism in supporting emotional control. The implementation includes the incorporation of the previous frame. The overall results are presented in Table 3.

Table 3: Comparison of methods with and without utilizing attention mechanism.

Method	M-LD ↓	M-LVD ↓	F-LD ↓	F-LVD ↓
w/o attention	2.11	1.35	1.97	1.29
w attention	<b>2.01</b>	<b>1.25</b>	<b>1.85</b>	<b>1.24</b>

The comparison results in Table 3 emphasize the performance distinction between methods that utilize the attention mechanism and those that do not. The measurements reveal that incorporating the attention mechanism for emotional control led to improved results of 2.01, 1.25, 1.85, and 1.24, respectively. These findings suggest that the inclusion of the attention mechanism effectively focuses on pertinent audio features based on emotion, resulting in enhanced outcomes for M-LD, M-LVD, F-LD, and F-LVD.

## 6 CONCLUSION

In conclusion, this paper introduces EAPC, a framework for generating realistic talking faces from audio and reference image input. We also propose the Dual-LSTM, which utilizes dual LSTM layers and incorporates skip connections from the prior audio frame to the current audio frame, thereby enhancing the temporal characteristics of our method. Additionally, the Dual-LSTM module employs the attention mechanism to support emotion control, effectively generating emotionally animated facial landmark frames. Qualitative results and our ablation study validate the effectiveness of our method, leading to the achievement of competitive results with state-of-the-art. This research opens up possibilities for more advanced and natural facial animation generation techniques in various applications, including video production, virtual avatars, and virtual reality experiences.

## ACKNOWLEDGEMENTS

This research is funded by the University of Science, VNU-HCM, Vietnam under grant number CNTT 2022-14.

## REFERENCES

- Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390.
- Chen, L., Li, Z., Maddox, R. K., Duan, Z., and Xu, C. (2018). Lip movements generation at a glance.
- Chen, L., Maddox, R. K., Duan, Z., and Xu, C. (2019). 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.
- Chung, J. S., Jamaludin, A., and Zisserman, A. (2017). You said that? *arXiv preprint arXiv:1705.02966*.
- Esikmez, S. E., Zhang, Y., and Duan, Z. (2021). Speech driven talking face generation from a single image and an emotion condition.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2018). Gans trained by a two time-scale update rule converge to a local nash equilibrium.
- Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.
- Ji, X., Zhou, H., Wang, K., Wu, Q., Wu, W., Xu, F., and Cao, X. (2022). Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10.
- Ji, X., Zhou, H., Wang, K., Wu, W., Loy, C. C., Cao, X., and Xu, F. (2021). Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14080–14089.
- Larkin, K. G. (2015). Structural similarity index simplified: Is there really a simpler concept at the heart of image quality measurement?
- Liu, X., Xu, Y., Wu, Q., Zhou, H., Wu, W., and Zhou, B. (2022). Semantic-aware implicit neural audio-driven video portrait generation. In *European Conference on Computer Vision*, pages 106–125. Springer.
- Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., et al. (2019). Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.
- Mittal, G. and Wang, B. (2020). Animating face using disentangled audio representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3290–3298.
- Sinha, S., Biswas, S., Yadav, R., and Bhowmick, B. (2022). Emotion-controllable generalized talking face generation.
- Song, Y., Zhu, J., Li, D., Wang, A., and Qi, H. (2019). Talking face generation by conditional recurrent adversarial network. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 919–925. International Joint Conferences on Artificial Intelligence Organization.
- Song, Y., Zhu, J., Li, D., Wang, X., and Qi, H. (2018). Talking face generation by conditional recurrent adversarial network. *arXiv preprint arXiv:1804.04786*.
- Suwajanakorn, S., Seitz, S. M., and Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph.*, 36(4).
- Thies, J., Elgharib, M., Tewari, A., Theobalt, C., and Nießner, M. (2020). Neural voice puppetry: Audio-driven facial reenactment.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Vougioukas, K., Petridis, S., and Pantic, M. (2019). Realistic speech-driven facial animation with gans.
- Wang, J., Qian, X., Zhang, M., Tan, R. T., and Li, H. (2023a). Seeing what you said: Talking face generation guided by a lip reading expert. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14653–14662.
- Wang, J., Zhao, K., Zhang, S., Zhang, Y., Shen, Y., Zhao, D., and Zhou, J. (2023b). Lipformer: High-fidelity and generalizable talking face generation with a pre-learned facial codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13844–13853.

- Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., and Loy, C. C. (2020). Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*.
- Wang, S., Li, L., Ding, Y., and Yu, X. (2022). 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.
- Wang, T.-C., Liu, M.-Y., Tao, A., Liu, G., Kautz, J., and Catanzaro, B. (2019). Few-shot video-to-video synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*.
- Zhou, H., Liu, Y., Liu, Z., Luo, P., and Wang, X. (2019). Talking face generation by adversarially disentangled audio-visual representation. In *AAAI Conference on Artificial Intelligence (AAAI)*.
- Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., and Li, D. (2020). Makeltalk. *ACM Transactions on Graphics*, 39(6):1–15.

