Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

SPECIAL ISSUE PAPER

Talking Face Generation with Lip and Identity Priors

Jiajie Wu[1] | **Frederick W. B. Li**[2] | **Gary K.L. Tam**[3] | **Bailin Yang**[1] | **Fangzhe Nan**[1] | **Jiahao Pan**[1]

[1]Department of Computer Science and Technology, Zhejiang Gongshang University, Hangzhou, China

[2]Department of Computer Science, University of Durham, Durham, United Kingdom

[3]Department of Computer Science, Swansea University, Swansea, United Kingdom

Correspondence
Bailin Yang, Department of Computer Science, Zhejiang Gongshang University,
Email: ybl@zjgsu.edu.cn

Funding Information
This research was supported by the Zhejiang Provincial Natural Science Foundation of China (Grant No. LD24F020003) and the National Natural Science Foundation of China (Grant No. 62172366).

Abstract

Speech-driven talking face video generation has attracted growing interest in recent research. While person-specific approaches yield high-fidelity results, they require extensive training data from each individual speaker. In contrast, general-purpose methods often struggle with accurate lip synchronization, identity preservation, and natural facial movements. To address these limitations, we propose a novel architecture that combines an alignment model with a rendering model. The rendering model synthesizes identity-consistent lip movements by leveraging facial landmarks derived from speech, a partially occluded target face, multi-reference lip features, and the input audio. Concurrently, the alignment model estimates optical flow using the occluded face and a static reference image, enabling precise alignment of facial poses and lip shapes. This collaborative design enhances the rendering process, resulting in more realistic and identity-preserving outputs. Extensive experiments demonstrate that our method significantly improves lip synchronization and identity retention, establishing a new benchmark in talking face video generation.

KEYWORDS
Talking Face Generation, Speech-Driven, Lip and Identity Priors

1 | INTRODUCTION

The objective of audio-driven talking face video generation is to create high-quality talking faces that are synchronized with spoken input. This task holds substantial value across various applications, including visual dubbing [1,2,3], digital assistants [4], and animated films [5]. Consequently, it has attracted significant attention from both industry and academia over the past decades.

Talking face video generation methods can primarily be categorized into two types: person-specific and general person methods. Person-specific approaches [4,6,7,8,9,10] excel at generating realistic talking face videos. However, they require distinct datasets for each individual and require re-training or fine-tuning for every new speaker, which considerably escalates training costs and complicates practical deployment. This limitation has spurred interest in generating talking face videos for general persons [1,2,3,11,12,13,14,15,16,17,18].

Two primary challenges arise in general talking face video generation: 1) accurately generating head and mouth movements that are synchronized with the audio, and 2) producing visually realistic results. To tackle these challenges, one approach involves using reconstruction-based models, which utilize an encoder-decoder architecture to generate talking videos in an end-to-end manner [2,12,13]. However, this method requires the model to learn pixel-level movements, which complicates the training process and hinders convergence.

An alternative strategy employs facial feature points as an intermediate structure to facilitate video generation [3,5,14,19,20,21,22]. This two-stage model approach allows the system to avoid directly learning low-level pixel appearances, thereby enabling it to focus more effectively on the facial motion trajectories of the individual. The initial stage model, IPLAP [14], first introduced the use of Transformers [23] to predict facial feature points, addressing the limitations of LSTM models in capturing long-term dependencies. To date, this model maintains state-of-the-art performance, and we utilize it to generate intermediate representations that are synchronized with the audio input.

In the subsequent stage, however, the lack of visual content in the input audio and intermediate landmarks presents a significant challenge. Generating realistic face videos while preserving identity information based solely on audio and intermediate landmarks remains a critical research issue that requires further exploration.

Existing methods, such as[3,5], rely on a single static reference image to provide visual appearance and identity information. However, the identity data derived from a single face is often insufficient, leading to unnatural distortions and stretching in the generated results. Consequently, these methods struggle to achieve satisfactory visual quality.

Other approaches, such as[19,24], utilize multiple reference images to offer richer details. However, when head movements dominate the frame, using multiple reference images without proper feature alignment can interfere with the generation process, despite the potential for enhanced identity representation. To address this issue, IPLAP[14] introduced an optical flow prediction network to align the features of the reference images, guiding the generation model towards improved outcomes.

Nevertheless, conventional optical flow prediction networks may experience reduced accuracy when there are significant differences in head motion magnitude and color between the reference and original images. This discrepancy can result in substantial artifacts in the generated guidance image, adversely affecting the performance of the guided model. Additionally, simply warping the reference image to create a guidance image can lead to the loss of original lip details and fails to provide audio-related guidance. As a result, the model struggles to accurately capture the lip features of the individual, hindering the generation of lip movements that are synchronized with the audio input.

To address the above challenges, we have developed a novel face generation framework that leverages multi-reference lip features, audio features, and aligned facial information to produce realistic face frames while preserving identity. We encompass a landmark-to-video generation network, which consists of an alignment model based on optical flow and a rendering model.

The alignment model predicts accurate optical flow by integrating inferred landmarks, occluded images, and reference images. It guides the rendering model by warping both the image features and those of the occluded original image. The rendering model is designed to accurately generate the lower part of the face, synchronized with the audio, using insights from the alignment model alongside an audio-aware cross-attention module.

Recognizing that the guidance information from the alignment model may yield distorted images rather than natural facial features, we have introduced a new module structure based on multiview lip perception. This structure enhances the rendering model's ability to accurately fit the lip features of the individual, moving beyond mere de-artifacting of the alignment model's outputs. Extensive experiments demonstrate that our method generates more realistic talking face videos and better preserves identity information and lip details compared to existing techniques. Our main contributions are:

- We propose a generation model framework comprising an alignment model and a rendering model, specifically designed to facilitate talking face generation guided by audio, multi-reference lip features, and aligned appearance information.
- We introduce an optical flow prediction network that utilises reference images and occluded original images to yield more accurately aligned reference images.
- We present a rendering generation network that is guided by a multi-reference lip module and an audio-aware cross-attention module, enhancing the realism of generated person features and lip details.

2 | RELATED WORK

2.1 | Speech-Driven Talking Face Generation

Existing methods for audio-driven talking face generation can be classified into two main categories: person-specific methods and general person methods. Person-specific techniques, leveraging 3D Morphable Models (3DMM)[25] and Neural Radiance Fields (NeRF)[26], have demonstrated the ability to synthesize high-fidelity talking face videos. For example, NVP[4] and FACIAL[10] first predict 3DMM expression parameters from audio and subsequently employ a neural rendering network to generate the video. Other works[6,7,9] achieve audio-driven talking face video generation by manipulating dynamic neural radiance fields based on audio input and rendering the corresponding facial images.

However, these methods require video data of the target speaker for retraining or fine-tuning, which can be impractical in many real-world scenarios. This limitation highlights the need for methods applicable to general faces. A body of literature addresses general person talking face generation, including approaches such as[2,12,16,17,18,21,27]. Generally, methods in this domain can be categorized into intermediate representation-based and reconstruction-based techniques[13].

Our proposed method effectively addresses these limitations by integrating multi-reference lip features and aligned facial information, enabling high-quality talking face generation without the need for extensive retraining on target-specific data.

2.2 | Reconstruction-Based Talking Face Generation

Reconstruction-based methods primarily employ an encoder-decoder architecture to generate talking face videos in an end-to-end fashion. For example, Wav2Lip[2] utilizes this structure to synthesize talking face videos with the assistance of a lip synchronization discriminator. Building on Wav2Lip, SyncTalkFace[12] introduces an audio-lip memory that provides additional visual information specifically for the mouth region. In contrast, PC-AVS[15] modularizes the talking face into distinct feature spaces for speech content, head pose, and identity, subsequently combining these features to generate the final video. TalkLip[13] employs a fine-tuned large lip-reading model, AVHubert, and integrates contrastive learning to enhance training effectiveness.

Despite their advantages, reconstruction-based methods are limited by the inherent structural design of the encoder-decoder framework. This requires the model to learn pixel-level motion, accounting for both the motion structure and rendering of the face, which complicates the training process. Furthermore, this complexity typically requires an increased number of parameters, resulting in higher training costs.

Our proposed method overcomes these challenges by utilizing a novel framework that effectively leverages multi-reference lip features and audio-guided alignment, streamlining the learning process while ensuring high fidelity in talking face generation.

2.3 | Intermediate Representation-Based Talking Face Generation

Intermediate representation-based methods employ two cascaded models that learn intermediate facial representations, such as facial landmarks or 3D meshes, for face synthesis from coherent speech input. Many audio-driven talking face generation methods[3,5,14,19,20,21,22,28] utilize facial landmarks as their intermediate representation, e.g., Suwajanakorn et al.[28] use a recurrent neural network (RNN) to map audio input to mouth landmarks, subsequently synthesizing high-quality mouth textures. Other studies[6,21,24] leverage long short-term memory (LSTM) models to learn the mapping between audio and landmark movements. MakeItTalk[5] integrates LSTM with a self-attention mechanism to predict landmark displacements from audio and generates face videos through a ResUnet network. IPLAP[14] employs a Transformer[23] model to learn the relationship between audio and facial landmarks, enhancing mapping accuracy.

Compared to reconstruction methods, intermediate representation-based techniques enable models to avoid directly learning low-level pixel appearances, allowing for a more focused understanding of facial and mouth motion trajectories[5]. Additionally, the separation of motion trajectory learning and rendering generation results in more compact and generalized models.

Our proposed method also consists of two stages: generation from audio to landmarks and rendering from landmarks to video. However, it introduces unique features that differentiate it from existing methods. First, we present a novel optical flow estimation network that effectively incorporates prior appearance information to guide the generation process. Second, our rendering generation network is enhanced by multi-reference lip features and audio features, ensuring the preservation of the person's identity and yielding more realistic lip details. Experimental results demonstrate that our method outperforms existing approaches.

3 | PROPOSED METHOD

Our objective is to generate a lip-synchronized talking face video from an audio sequence and an initial input video. The proposed method consists of two main stages, as illustrated in Figure 1.

In the first stage, we utilize the IPLAP speech-to-facial landmark model to predict facial landmarks from the input speech sequence. The model takes as input the speech signal, reference facial landmarks, and the upper half of the actual facial landmarks. It outputs the positions of the lip and lower jaw landmarks. Leveraging a Transformer architecture, the model captures long-term dependencies in the audio and facial landmark data, thereby enhancing prediction accuracy through the integration of reference facial landmarks.

The second stage encompasses a two-part generation model: the alignment model and the rendering model. Initially, the rendering model extracts lip information corresponding to the person's identity by utilizing reference lip features, the occluded

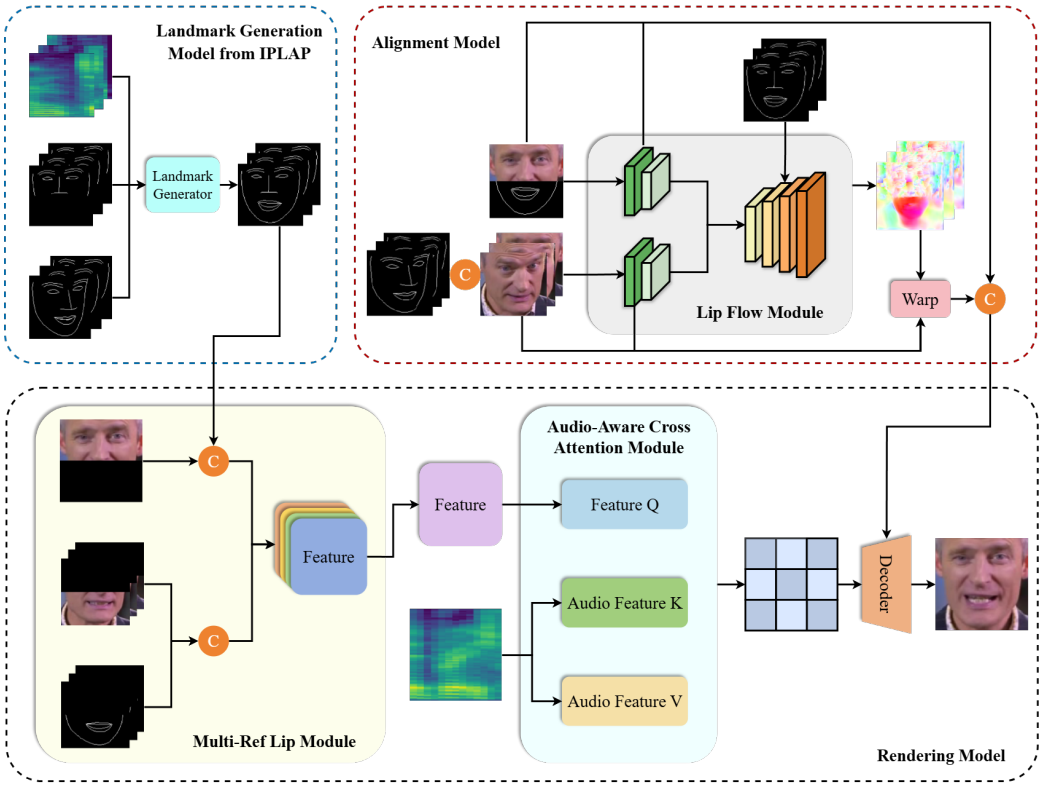

FIGURE 1 Overview of the Architecture of Our Proposed Method

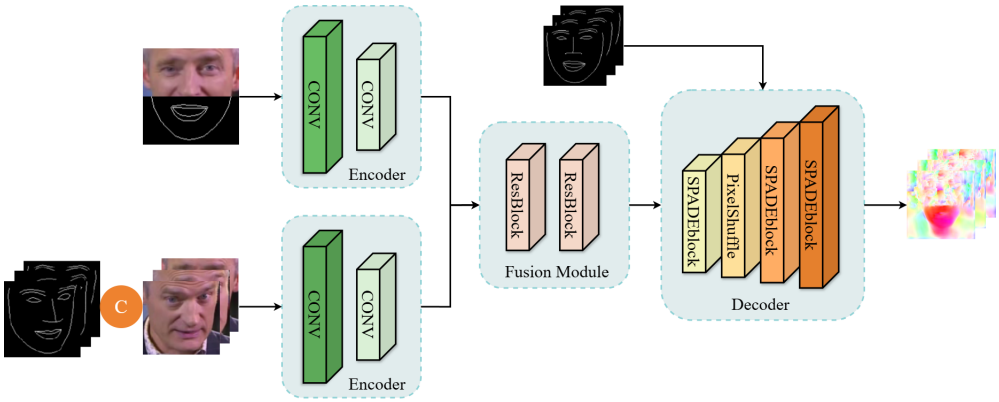

FIGURE 2 Lip Flow Module

real face, and audio features. Subsequently, the alignment model generates optical flow based on the predicted facial landmarks, reference face, and occluded real face. These optical flow features are used to warp the image data. Finally, the warped features, combined with the reference image, guide the rendering model to produce the synthesized video.

Our approach does not only ensure accurate synchronization of lip movements with the audio but also preserves the identity and realism of the generated talking face.

3.1 | Audio-To-Landmark Generation

To predict facial landmarks from speech, we utilize the IPLAP[14] speech-to-facial landmark generation model. This model takes as inputs the reference facial landmarks, mel-spectrogram, and the upper half of the actual facial landmarks, employing a Transformer to capture long-sequence relationships and feature dependencies. The output is the location of the lip and lower jaw landmarks.

Specifically, we begin by using a facial landmark detector to extract 2D facial landmarks from the video. The model incorporates three input branches: the reference facial landmark set $\{l_i^r \in \mathbb{R}^{2\times n_r}\}_{i=1}^{N_l}$ from the original video, the mel-spectrogram sequence $\{m_t \in \mathbb{R}^{h\times w}\}_{t=1}^{T}$ corresponding to the video frames, and the upper half of the real label facial pose landmarks $\{l_t^p \in \mathbb{R}^{2\times n_p}\}_{t=1}^{T}$. Here, n_r and n_p denote the total number of facial landmarks and facial pose landmarks, respectively. The model extracts prior knowledge from the reference facial landmark set l_i^r, combines it with the speech features and facial pose landmark features, and incorporates positional encoding to accurately predict the locations of the lip and lower jaw landmarks synchronized with the audio.

3.2 | Landmark-To-Video Generation

To enhance the rendering model's pixel feature guidance, we developed a generation model comprising two key components: the alignment model M_a and the rendering model M_r. Initially, we combine the predicted lip and lower jaw landmarks with the real upper half landmarks to create a comprehensive set of facial landmarks $\{L_i \in \mathbb{R}^{3\times H\times W}\}_{i=t-k}^{t+k}$ surrounding the t-th frame of the video, incorporating $2k+1$ frames.

Next, we concatenate the occluded real face image $I_t^m \in \mathbb{R}^{3\times H\times W}$ with the facial landmarks along the channel dimension. We also extract multiple frames of lip reference images $\{I_i^l \in \mathbb{R}^{3\times H\times W}\}_{i=1}^{N}$ and their corresponding lip landmark points $\{L_i^l \in \mathbb{R}^{3\times H\times W}\}_{i=1}^{N}$ from the original video, which are then fed into the multi-reference lip module of the rendering model. This module fuses the features with real lip information, enhancing the accuracy of mouth shape generation.

To further refine the realism of the facial appearances generated by the rendering model, we input the occluded real face I_t^m, multiple reference face images $\{I_i^r \in \mathbb{R}^{3\times H\times W}\}_{i=1}^{N}$, and their corresponding facial landmarks $\{L_i^r \in \mathbb{R}^{3\times H\times W}\}_{i=1}^{N}$ into the alignment model. This model computes the flow fields $\{F_i \in \mathbb{R}^{2\times H\times W}\}_{i=1}^{N}$, which are then used to warp the reference images and features to align with the target head pose and expression.

Finally, the rendering model synthesizes the target face image by leveraging the audio features, the warped and aligned reference images and features, the occluded real face image, and the multi-reference mouth images. This comprehensive approach ensures that the generated video maintains high fidelity in both lip synchronization and facial realism.

3.2.1 | Alignment Model

The alignment model M_a is primarily focused on lip flow prediction, being essential for generating accurate facial movements in response to audio input. As in Figure 2, the model architecture consists of two encoders E_r and E_m that extract features from the reference image I^r and the occluded real face image I_t^m, respectively.

First, we concatenate the features extracted from both encoders along the channel dimension, creating a richer feature representation. This combined feature set is then input into a feature fusion module built with Resblocks, which enhances the model's capacity to learn complex patterns in the data.

To guide the prediction of the flow field more effectively, we incorporate facial landmarks L into the alignment model using the SPADE[29] layer. This allows the model to leverage spatial information from the landmarks to improve the accuracy of flow field predictions. The functional representation of the alignment model is:

$$F_i = M_a(I_t^m, L_i^r, I_i^r, L_{t-k:t+k}) \quad i = 1, .., N$$

where F_i denotes the flow field that guides the warping of reference images and their corresponding features. By utilizing the appearance information from both the reference image and the occluded real face image, alongside the facial landmarks, the alignment model predicts a more accurate flow field, facilitating the precise warping of features for the rendering model.

Additionally, to aggregate visual features warped by the flow fields from multiple reference images, the alignment model includes an output layer that predicts a spatial weight $w_i \in \mathbb{R}^{H\times W}$ for each reference image I_i^r. The aggregated aligned image is

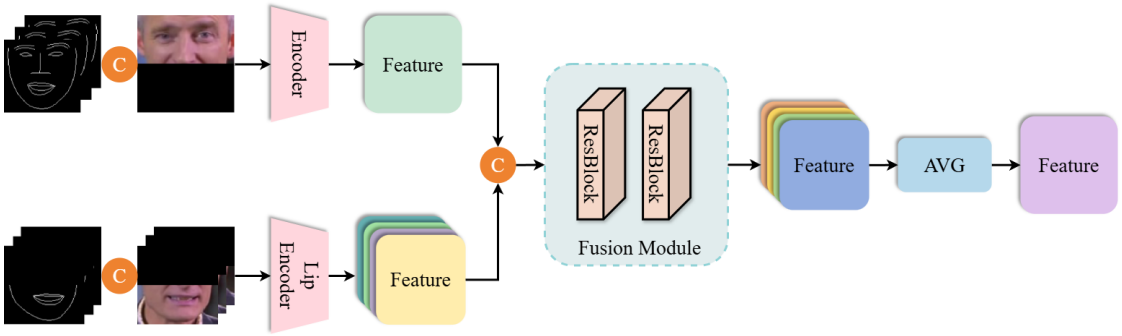

FIGURE 3 Multi-Ref Lip Module

computed as:

$$\overline{I^r} = \frac{\sum_{i=1}^{N} w_i F_i(I_i^r)}{\sum_{i=1}^{N} w_i}$$

where $\overline{I^r}$ represents the aggregated aligned image. This process ensures that the most relevant visual information from multiple references is effectively combined, enhancing the overall quality of the generated output.

To guide the high-dimensional features of the rendering model, we warp and align the features $f_i^{r_1} \in \mathbb{R}^{c_1 \times h_1 \times w_1}$ and $f_i^{r_2} \in \mathbb{R}^{c_2 \times h_2 \times w_2}$ extracted by the encoder E_r at two spatial resolutions. These aligned features calculated by:

$$f_i^{r_1}, f_i^{r_2} = E_r(L_i^r, I_i^r) \quad i = 1, 2, \ldots, N$$

$$\overline{f^{r_x}} = \frac{\sum_{i=1}^{N} w_i F_i(f_i^{r_x})}{\sum_{i=1}^{N} w_i} \quad x = 1, 2$$

where $\overline{f^{r_x}}$ represents the aggregated aligned high-dimensional features at resolutions $x = 1$ and $x = 2$.

To mitigate the effects of color discrepancies and artifacts caused by large head motions, we concatenate the feature of the occluded real face image with the warped features along the channel dimension. This concatenation provides a robust guide for the rendering model's generation process. The operations are:

$$f^{m_1}, f^{m_2} = E_m(I_t^m)$$

$$I^c, f^{c_x} = \text{concat}[I_t^m, \overline{I^r}], \text{concat}[f^{m_x}, \overline{f^{r_x}}] \quad x = 1, 2$$

where f^{m_1} and f^{m_2} are the features extracted from the occluded real face image by the encoder E_m at two different spatial resolutions. The jointly guiding image I^c and the features f^{c_x} are produced after channel concatenation.

Finally, the concatenated outputs I^c and f^{c_x} are fed through the SPADE layer into the decoder of the rendering model. This process ensures that the generated facial animations are realistic and accurately synchronized with the audio input, addressing key challenges in talking face generation.

3.2.2 | Rendering Model

The rendering model M_r is composed of two primary components: the multi-reference lip module and the audio-aware cross-attention module. This model is designed to generate high-quality facial animations that accurately synchronize with audio input, addressing key challenges in talking face generation.

As in Figure 3, the process begins by extracting features from the occluded real face image $f^m \in \mathbb{R}^{c \times h \times w}$ and the reference lip features $f_i^l \in \mathbb{R}^{c \times h \times w}$ using two distinct encoders. These features are then concatenated along the channel dimension and passed into a feature fusion module F constructed from Resblocks, which facilitates the combination of information from both sources. The resulting fused features $f_i \in \mathbb{R}^{c \times h \times w}$ are computed as:

$$f_i = F\left(\text{concat}[f^m, f_i^l]\right) \quad i = 1, 2, \ldots, N$$

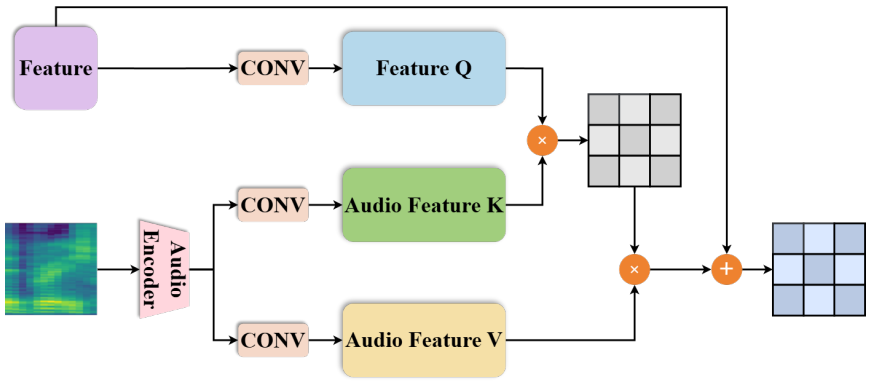

FIGURE 4 Audio-Aware Cross-Attention Module

To obtain the final multi-reference lip feature that captures the speaker's identity, we take the average of the set of fused features:

$$\bar{f} = \text{AVG}(f_i) \quad i = 1, 2, \ldots, N$$

where $\bar{f}$ represents the aggregated multi-reference lip feature.

Next, we integrate the audio-aware cross-attention module as shown in Figure 4, which plays a crucial role in synchronizing the visual features with the corresponding audio. This is accomplished by fusing the mel-spectrogram m_t of the video frame with the aggregated image features $\bar{f}$. The audio features for the t-th frame are obtained using the audio encoder E_a:

$$f_t^a = E_a(m_t)$$

The cross-attention mechanism is implemented using:

$$f_g = \text{Softmax}\left((W_q\bar{f})(W_k f_t^a)^T\right) \times (W_v f_t^a)$$

where W_q, W_k, and W_v are three distinct 1×1 convolutional layers that project the features into the appropriate dimensions for attention computation. Here, f_g represents the fused features that guide the generation of the mouth shape synchronized with the audio.

Finally, the rendering model generates the target image $\hat{I}_t$ by combining the attention-fused features f_g with the features I^c and f^{c_x} obtained from the alignment model. This is achieved through a decoder that employs SPADE and PixelShuffle[30] layers to produce the final output. The overall function of the rendering model is expressed as:

$$\hat{I}_t = M_r\left(I_t^m, I_i^l, L_{t-k:t+k}, I^c, f^{c_1}, f^{c_2}, m_t\right) \quad i = 1, \ldots, N$$

where $\hat{I}_t$ is the synthesized image for the t-th frame, ensuring that the generated facial animations are not only visually coherent but also accurately synchronized with the input audio, thereby effectively addressing the challenges associated with realistic talking face generation.

3.2.3 | Loss Function for Landmark-To-Video

Our loss function for the alignment model is based on the structure-aware perceptual loss L_w, which leverages the VGG-19 network to enhance the accuracy of flow fields and weights. This loss function compares the ground truth image I_t with the warped and aligned image $\overline{I^r}$:

$$L_w = \sum_i \|\phi_i(\overline{I^r}) - \phi_i(I_t)\|_1$$

where ϕ_i denotes the activation output from the i-th layer of the VGG-19 network, allowing us to capture perceptual differences between the generated and target images at multiple levels of abstraction.

For the rendering model, we employ a combination of reconstruction loss L_r and style loss L_s. Both losses follow a structure similar to L_w and are used to compare the generated image $\hat{I}_t$ with the ground truth I_t, thereby improving the quality of the

TABLE 1 Quantitative comparison of different methods on LRW and LRS2 datasets.

Method	Dataset	PSNR↑	SSIM↑	LPIPS↓	FID↓	LipLMD↓	LSE-D↓
Wav2lip	LRW	26.87	0.8897	0.0760	26.29	0.996	7.91
PC-AVS		19.64	0.6001	0.3008	67.57	3.707	**7.78**
EmoGen		19.43	0.5181	0.2424	46.54	4.800	8.74
IPLAP		27.08	0.8903	0.0583	23.59	1.251	8.51
DiffDub		26.27	0.8541	0.1099	27.53	3.095	8.30
Ours		**30.88**	**0.9327**	**0.0356**	**15.05**	**0.985**	8.14
Wav2lip	LRS2	25.82	0.8581	0.1027	18.22	1.414	**6.76**
PC-AVS		18.51	0.5558	0.3626	50.76	8.412	6.78
EmoGen		19.99	0.5115	0.2766	48.11	6.451	7.91
IPLAP		28.28	0.8792	0.0617	15.61	1.538	7.11
DiffDub		25.99	0.8056	0.1311	37.60	2.186	8.08
Ours w/o Ali		25.42	0.8503	0.0978	23.31	1.574	-
Ours w/o Lip		31.22	0.9171	0.0364	11.64	1.185	-
Ours w/o Aud		32.70	0.9245	0.0370	10.41	1.202	7.65
Ours		**32.93**	**0.9261**	**0.0355**	**9.89**	**1.128**	6.97

generated results:

$$L_r = \sum_i \|\phi_i(\hat{I}_t) - \phi_i(I_t)\|_1$$

$$L_s = \sum_i \|G_i^{\phi}(\hat{I}_t) - G_i^{\phi}(I_t)\|_1$$

where G_i^{ϕ} represents the Gram matrix derived from the activation layer output ϕ_i, which captures style information by comparing the correlations between feature maps.

To further enhance the realism of the rendered images, we also incorporate the patch GAN loss L_g and feature matching loss L_f, inspired by the pix2pixHD framework[31]. The overall loss function L for our model is thus defined as:

$$L = \lambda_w L_w + \lambda_r L_r + \lambda_s L_s + \lambda_g L_g + \lambda_f L_f$$

where λ_w, λ_r, λ_s, λ_g, and λ_f are the weighting parameters that balance the contributions of each loss component during training. For our experiments, we set the parameters as follows: $\lambda_w = 2.5$, $\lambda_r = 4$, $\lambda_s = 1000$, $\lambda_g = 0.25$, and $\lambda_f = 2.5$.

This comprehensive loss function effectively guides the model to generate high-fidelity talking face videos that accurately synchronize with audio while preserving the identity and realism of the speaker's appearance.

4 | EXPERIMENTS

4.1 | Experimental Settings

Implementation Details: During inference, we adhere to the settings outlined in[14] for processing video frames and audio. Specifically, video frames are center-cropped and resized to 128×128 pixels. The audio is processed using a window size of 800 samples and a hop size of 200 samples, resulting in an 80×16 mel-spectrogram at a sampling rate of 16 kHz. This configuration allows for effective extraction of audio features while maintaining temporal coherence in the spectrogram representation. This choice also ensures compatibility and a fair comparison with existing methods.

In the alignment model, *ref* – *N* is set to 3 during training to enhance accuracy through multiple reference frames, while during inference, it is adjusted to one-fifth of the video length for efficiency. The rendering model uses *ref* – *lip* set to 5, leveraging diverse lip features for better identity representation. All experiments are run on 4 L40 GPUs for adequate computational power.

Datasets: We train our method on the LRS2 dataset[32], which comprises a rich collection of video clips. For evaluation and comparison with existing methods, we utilize the test sets of two public datasets: LRS2 and LRW[33]. The LRW dataset is specifically designed for audio-visual word classification, containing 500 distinct words, each represented by a video of 29 frames (approximately 1 second) synchronized with audio. The LRS2 dataset consists of 48,164 video clips, including 45,839 for training, 1,082 for validation, and 1,243 for testing. We follow IPLAP[14] in our experimental setup and sample 45 videos from the test set of each dataset to ensure a robust evaluation of our method's performance.

Metrics: For quality assessment of the generated images, we employ several metrics: PSNR[12] and SSIM[34] measure the similarity between generated images and real images, providing insights into fidelity and structural similarity. Also, LPIPS[35] and FID[36] evaluate feature similarity, offering a detailed understanding of perceptual differences.

To assess lip synchronization capabilities, we utilize LMD[3] and LSE-D[2]. LMD quantifies the distance between generated and real lip landmarks, providing a direct measure of synchronization accuracy. LSE-D, on the other hand, employs a pre-trained SyncNet[37] to quantify lip synchronization performance, ensuring comprehensive evaluation of our method's efficacy.

Comparison: We compare our method against four state-of-the-art talking face generation techniques: EmoGen, PC-AVS, Wav2Lip, IPLAP, and DiffDub. EmoGen[38] generates lip-synced videos while allowing for emotion control through the injection of emotion labels. PC-AVS[15] enables pose-controllable talking face generation using modular audiovisual representations. Wav2Lip[2] employs an adversarially trained encoder-decoder model to create synchronized talking face videos. IPLAP[14] focuses on generating talking face videos by driving 2D landmarks based on input audio. Finally, DiffDub[17] generates audio-synchronized mouth shapes via an inpainting renderer based on a diffusion auto-encoder.

This comprehensive experimental setup, including carefully selected metrics and comparison methods, provides a robust framework for evaluating the effectiveness of our method in generating high-quality talking face videos.

4.2 | Experimental Results

Quantitative Results: Table 1 demonstrates that our method significantly improves image quality metrics, including PSNR, SSIM, and FID. These improvements indicate that our approach generates more detailed lip features while effectively preserving the identity of the individual. Specifically, higher PSNR values reflect enhanced signal fidelity, while SSIM scores suggest better structural similarity to the ground truth. Our method also excels in lip synchronization, achieving the best performance in the LMD metric, which quantifies the distance between generated and real lip landmarks.

However, we observe that our model slightly lags behind Wav2Lip and PC-AVS in audiovisual synchronization, as measured by the LSE-D metric. We attribute this to subtle temporal jitters in the lip landmarks produced by the LandmarkGenerator, which can negatively affect the synchronization accuracy.

Visualization of Generated Images: Figure 5 illustrates that our generated images are visually closer to the real ground truth labels compared to competing methods, exhibiting fewer artifacts. Moreover, our approach excels in restoring fine details of the person's lips, further enhancing the realism of the generated outputs. This ability to capture subtle lip movements highlights the effectiveness of our alignment and rendering strategies in producing high-quality talking face videos.

4.3 | Ablation Study on the Alignment Model

To assess the effectiveness of the alignment model, we conducted an ablation study comparing the full model with a variant that excludes the alignment module, as presented in Table 1. Our findings indicate a notable decrease in image quality metrics when the alignment model is removed. This decline underscores the critical role of the alignment model in preserving the subject's identity and ensuring high-quality image generation. Specifically, the metrics such as PSNR and SSIM, which quantify fidelity and structural similarity, showed significant reductions, affirming that the alignment model effectively enhances the overall performance of our system. This confirms that integrating the alignment module is essential for achieving accurate and realistic talking face videos.

FIGURE 5 Qualitative Comparison of Different Methods on LRS2 Dataset

4.4 | Ablation Study on the Multi-Ref Lip Module

To validate the effectiveness of the multi-Ref lip reference module, we conducted an ablation study by removing this module from the rendering model, as detailed in Table 1. The results reveal a marked decline in both image quality metrics and the lip synchronization metric (LipLMD) when the multi-Ref lip reference module is excluded. This decrease highlights the critical contribution of the multi-Ref lip module in generating realistic lip details. Specifically, the reduce in performance of PSNR and LipLMD indicate that the module does not only enhance the fidelity of the generated images but also significantly improves the accuracy of lip movements. These findings confirm that the multi-Ref lip reference module is essential for producing high-quality, synchronized talking face videos.

4.5 | Ablation Study on the Audio-Aware Cross Attention Module

To validate the effectiveness of the audio-aware cross-attention module, we performed an ablation study by removing this module from the rendering model, as illustrated in Table 1. The results indicate that incorporating the audio-aware cross-attention module leads to significant improvements in image quality metrics, such as PSNR and SSIM, as well as in lip synchronization metrics, including LipLMD and LSE-D. The enhancements in these metrics suggest that the audio-aware cross-attention module effectively aligns audio features with visual content, resulting in clearer image details and more accurate lip movements. The mathematical foundation behind this improvement lies in the attention mechanism, which allows the model to focus on relevant

audio-visual correlations, thus enhancing both the quality of generated images and the synchronization of lip movements. These findings confirm the critical role of the audio-aware cross-attention module in achieving high-quality, synchronized talking face videos.

4.6 | Impact of Number of Reference Images and Lips

To assess the impact of the number of reference images and lips on model performance, we conducted ablation experiments, as shown in Tables 2 and 3. Table 2 reveals that the model achieves optimal performance with 25 reference images, reaching a PSNR of 32.93 and an SSIM of 0.9261, while the FID drops to 9.89, indicating improved image quality. Similarly, Table 3 demonstrates that using 5 reference lips yields the best results.

We further observe that: (1) performance improvements tend to plateau around 25 reference images and 5 reference lips, as supported by both our empirical visual observations and quantitative ablation study results; and (2) increasing the number of references results in significantly longer inference times and higher GPU memory usage. Taken together, these lead us to empirically select 25 reference images and 5 reference lips, which offer an effective balance between generation quality and computational efficiency.

TABLE 2 Results of different reference images.

Ref Num	PSNR↑	SSIM↑	LPIPS↓	FID↓
N=1	27.86	0.8718	0.0917	20.16
N=5	29.32	0.8912	0.0588	13.15
N=10	31.71	0.9159	0.0419	10.64
N=25	**32.93**	**0.9261**	**0.0355**	**9.89**

TABLE 3 Results of different lip numbers.

Lip Num	PSNR↑	SSIM↑	LPIPS↓	FID↓
N=1	31.88	0.9191	0.0362	10.06
N=3	32.59	0.9251	0.0359	9.89
N=5	**32.93**	**0.9261**	**0.0355**	**9.89**

5 | CONCLUSION

In this paper, we proposed a comprehensive model framework for talking face generation that integrates an alignment model and a rendering model, utilizing multi-Reference lip features and aligned appearance guidance. Our approach begins with the multi-Ref lip feature extraction module, which effectively fuses reference lip features with the upper facial features of the target individual. We then employ an audio-aware cross-attention module to merge audio features with facial features, significantly enhancing lip synchronization. Finally, the optical flow alignment model warps the reference person's image, guiding the rendering process to produce high-quality outputs. Experimental results demonstrate that our method surpasses existing approaches in preserving person identity, enhancing lip detail, and ensuring accurate mouth synchronization. For future work, we aim to explore the integration of more complex audio-visual cues and investigate real-time applications to further enhance the robustness and applicability of our model.

ACKNOWLEDGMENTS

This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LD24F020003. The authors also acknowledge the support from the NSFC (National Natural Science Foundation of China) under Grant No. 62172366. Gary Tam is supported by the Royal Society grant IEC/NSFC/211159. For the purpose of Open Access the author has applied a CC BY copyright licence to any Author Accepted Manuscript version arising from this submission.

REFERENCES

1. KR P, Mukhopadhyay R, Philip J, Jha A, Namboodiri V, Jawahar C. Towards automatic face-to-face translation. In: 2019:1428–1436.

2. Prajwal K, Mukhopadhyay R, Namboodiri VP, Jawahar C. A lip sync expert is all you need for speech to lip generation in the wild. In: 2020:484–492.
3. Xie T, Liao L, Bi C, et al. Towards realistic visual dubbing with heterogeneous sources. In: 2021:1739–1747.
4. Thies J, Elgharib M, Tewari A, Theobalt C, Nießner M. Neural voice puppetry: Audio-driven facial reenactment. In: Springer. 2020:716–731.
5. Zhou Y, Han X, Shechtman E, Echevarria J, Kalogerakis E, Li D. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG).* 2020;39(6):1–15.
6. Guo Y, Chen K, Liang S, Liu YJ, Bao H, Zhang J. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In: 2021:5784–5794.
7. Liu X, Xu Y, Wu Q, Zhou H, Wu W, Zhou B. Semantic-aware implicit neural audio-driven video portrait generation. In: Springer. 2022:106–125.
8. Lu Y, Chai J, Cao X. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (ToG).* 2021;40(6):1–17.
9. Shen S, Li W, Zhu Z, Duan Y, Zhou J, Lu J. Learning dynamic facial radiance fields for few-shot talking head synthesis. In: Springer. 2022:666–682.
10. Zhang C, Zhao Y, Huang Y, et al. Facial: Synthesizing dynamic talking face with implicit attribute learning. In: 2021:3867–3876.
11. Huang R, Zhong W, Li G. Audio-driven talking head generation with transformer and 3d morphable model. In: 2022:7035–7039.
12. Park SJ, Kim M, Hong J, Choi J, Ro YM. Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory. In: . 36. 2022:2062–2070.
13. Wang J, Qian X, Zhang M, Tan RT, Li H. Seeing what you said: Talking face generation guided by a lip reading expert. In: 2023:14653–14662.
14. Zhong W, Fang C, Cai Y, et al. Identity-preserving talking face generation with landmark and appearance priors. In: 2023:9729–9738.
15. Zhou H, Sun Y, Wu W, Loy CC, Wang X, Liu Z. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: 2021:4176–4186.
16. Shen S, Zhao W, Meng Z, et al. DiffTalk: Crafting Diffusion Models for Generalized Audio-Driven Portraits Animation. In: 2023.
17. Liu T, Du C, Fan S, Chen F, Yu K. DiffDub: Person-Generic Visual Dubbing Using Inpainting Renderer with Diffusion Auto-Encoder. In: IEEE. 2024:3630–3634.
18. Cui J, Li H, Yao Y, et al. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. *arXiv preprint arXiv:2410.07718.* 2024.
19. Biswas S, Sinha S, Das D, Bhowmick B. Realistic talking face animation with speech-induced head motion. In: 2021:1–9.
20. Chen L, Cui G, Liu C, et al. Talking-head generation with rhythmic head motion. In: Springer. 2020:35–51.
21. Chen L, Maddox RK, Duan Z, Xu C. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: 2019:7832–7841.
22. Ji X, Zhou H, Wang K, et al. Audio-driven emotional video portraits. In: 2021:14080–14089.
23. Vaswani A. Attention is all you need. *Advances in Neural Information Processing Systems.* 2017.
24. Chung JS, Jamaludin A, Zisserman A. You said that?. *arXiv preprint arXiv:1705.02966.* 2017.
25. Blanz V, Vetter T. A morphable model for the synthesis of 3D faces. In: , , , 2023:157–164.
26. Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM.* 2021;65(1):99–106.
27. Ji X, Zhou H, Wang K, et al. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In: 2022:1–10.
28. Suwajanakorn S, Seitz SM, Kemelmacher-Shlizerman I. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG).* 2017;36(4):1–13.
29. Park T, Liu MY, Wang TC, Zhu JY. Semantic image synthesis with spatially-adaptive normalization. In: 2019:2337–2346.
30. Shi W, Caballero J, Huszár F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: 2016:1874–1883.
31. Wang TC, Liu MY, Zhu JY, Tao A, Kautz J, Catanzaro B. High-resolution image synthesis and semantic manipulation with conditional gans. In: 2018:8798–8807.
32. Afouras T, Chung JS, Senior A, Vinyals O, Zisserman A. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence.* 2018;44(12):8717–8727.
33. Chung JS, Zisserman A. Lip reading in the wild. In: Springer. 2017:87–103.
34. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing.* 2004;13(4):600–612.
35. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: 2018:586–595.
36. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems.* 2017;30.
37. Chung JS, Zisserman A. Out of time: automated lip sync in the wild. In: Springer. 2017:251–263.
38. Goyal S, Bhagat S, Uppal S, et al. Emotionally enhanced talking face generation. In: 2023:81–90.

AUTHOR BIOGRAPHY

Jiajie Wu is currently pursuing his studies at Zhejiang Gongshang University. His main research interests lie in the field of computer vision, particularly in talking face generation. He explores how techniques in artificial intelligence programming, computer animation, and computer graphics can be leveraged to improve talking face synthesis.

Frederick W. B. Li received a B.A. and an M.Phil. degree from Hong Kong Polytechnic University, and a Ph.D. degree from the City University of Hong Kong. He is currently an Associate Professor at Durham University, researching computer graphics, deep learning, collaborative virtual environments, and educational technologies. He is also an Editorial Board Member of Virtual Reality & Intelligent Hardware. He chaired conferences such as ISVC and ICWL.

Gary K.L. Tam is a Senior Lecturer in the Department of Computer Science at Swansea University. He holds MPhil and Ph.D. degrees from City University of Hong Kong and Durham University. His recent research interests focus on 3D geometry and computer vision. Gary has served as a Guest Editor for special issues of International Journal of Computer Vision (2017), Computers (2018, 2019), and Applied Sciences (2023, 2024). He is also an Associate Editor for AI Communications since 2024.

Bai-Lin Yang received his Bachelor's and Ph.D. degrees from Dept. Computer Science, Hangzhou Dianzi University in 2003 and Zhejiang University in 2007, respectively. Now, he is a faculty member of Zhejiang Gongshang University. Yang's research interests include web graphics, realtime rendering and mobile game.

Fangzhe Nan, a doctoral candidate at Zhejiang Gongshang University, received the B.S. and master's degrees form Xinjiang University, Urumqi, China, in 2017 and 2020, respectively.Her research interests include computer vision, video and image processing, and point cloud compression.

Jiahao Pan is currently pursuing a Master's degree at Zhejiang Gongshang University. His research interests include computer graphics, deep learning, facial animation, and talking faces. Currently, he focuses on employing artificial intelligence technology to enhance the performance of speech-driven 3D facial animation.