High-Fidelity and Lip-Synced Talking Face Synthesis via Landmark-Based Diffusion Model

Weizhi Zhong, Junfan Lin, Peixin Chen, Feng Gao, *Member, IEEE*, Liang Lin, *Fellow, IEEE*, and Guanbin Li, *Member, IEEE*

***Abstract*—Audio-driven talking face video generation has attracted increasing attention due to its huge industrial potential. Some previous methods focus on learning a direct mapping from audio to visual content. Despite progress, they often struggle with the ambiguity of the mapping process, leading to flawed results. An alternative strategy involves facial structural representations (e.g., facial landmarks) as intermediaries. This multi-stage approach better preserves the appearance details but suffers from error accumulation due to the independent optimization of different stages. Moreover, most previous methods rely on generative adversarial networks, prone to training instability and mode collapse. To address these challenges, our study proposes a novel landmark-based diffusion model for talking face generation, which leverages facial landmarks as intermediate representations while enabling end-to-end optimization. Specifically, we first establish the less ambiguous mapping from audio to landmark motion of lip and jaw. Then, we introduce an innovative conditioning module called TalkFormer to align the synthesized motion with the motion represented by landmarks via differentiable cross-attention, which enables end-to-end optimization for improved lip synchronization. Besides, TalkFormer employs implicit feature warping to align the reference image features with the target motion for preserving more appearance details. Extensive experiments demonstrate that our approach can synthesize high-fidelity and lip-synced talking face videos, preserving more subject appearance details from the reference image.**

***Index Terms*—Talking face generation, landmark-based, high-fidelity lip synchronization, diffusion model.**

Received 23 May 2024; revised 4 August 2025; accepted 4 March 2026. Date of publication 26 March 2026; date of current version 31 March 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62322608 and Grant 62506180, in part by China Postdoctoral Science Foundation under Grant 2025M771522, and in part by Huawei's AI Hundred Schools Program and was carried out using Huawei Ascend AI Technology Stack. The associate editor coordinating the review of this article and approving it for publication was Dr. Shengcai Liao. *(Corresponding authors: Feng Gao; Guanbin Li.)*

Weizhi Zhong is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510275, China, and also with the Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: zhongwzh5@mail2.sysu.edu.cn).

Junfan Lin is with the Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: linjf@pcl.ac.cn).

Peixin Chen is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510275, China (e-mail: chenpx28@mail2.sysu.edu.cn).

Feng Gao is with the Peking University, Beijing 100091, China (e-mail: gaof@pku.edu.cn).

Liang Lin and Guanbin Li are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China, also with the Peng Cheng Laboratory, Shenzhen 518000, China, and also with Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou 510006, China (e-mail: linliang@ieee.org; liguanbin@mail.sysu.edu.cn).

Digital Object Identifier 10.1109/TIP.2026.3676251

I. Introduction

AUDIO-DRIVEN talking face video generation, a challenging task of cross-modal synthesis, aims to create talking videos with lip movements that are accurately synchronized with the input audio. This task has attracted increasing interest from the research community due to its broad applications, such as visual dubbing [1], virtual avatars [2], and digital humans [3]. To generate high-fidelity talking face videos, a prevalent manner is to gather data of the target individual to learn a person-specific model [4], [5], [6], [7]. While effective, these person-specific methods are hindered by the costly data collection and extensive training process. In contrast, person-generic methods can generalize to unseen subjects without further training. However, these methods often grapple with challenges in maintaining the appearance details of subjects as well as lip-audio synchronization. In this study, we aim to develop a person-generic talking face generation framework that generates high-fidelity facial details for general subjects while ensuring accurate lip-audio synchronization.

To achieve faithful talking face generation, two critical issues need to be considered: the high fidelity of the subject appearance details and the synchronization of lip movement with the audio input. To generate lip-synced talking videos, prior methods [8], [9], [10] directly model the mapping between audio signal and visual content. However, such audio-visual mapping is often uncertain and ambiguous, as one phonetic unit can potentially match various visual forms due to the diversity in illumination, emotion, and appearance. This often leads to flawed results and the loss of details in subject appearance. To ease the ambiguity for improved fidelity of the subject appearance, another line of works [1], [11], [12], [13], [14], [15] instead first establish a correlation between the audio signals and intermediate structural representation such as facial landmarks or 3D Morphable Model (3DMM) [16], which primarily reflect the motion information and facilitate the less ambiguous mapping from audio to motion. Then, another stage converts these motion representations into realistic facial images. However, a notable drawback of these approaches is the isolated training of different stages, potentially resulting in inaccuracies in lip-audio synchronization stemming from errors in the pre-estimated structural representations. Additionally, prior methods [8], [9], [11], [15] frequently rely on either misaligned reference images or images

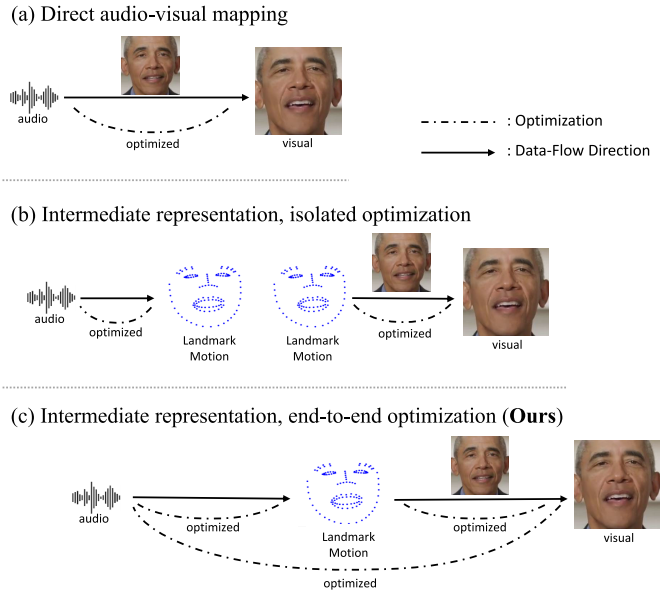

Fig. 1. An illustration of three strategies to learn the audio-visual relationship for talking face generation. (a) Previous methods optimize a network to learn the direct audio-visual mapping, resulting in flawed results. (b) Previous methods optimize a network to learn the mapping from audio to landmark motion, then separately optimize another mapping from the motion representation to a realistic face, suffering from the inaccuracies of pre-estimated intermediate representation. (c) Our method leverages facial landmarks as intermediate representation while enabling end-to-end optimization to reduce the errors accumulation resulting from pre-estimated landmark inaccuracies.

warped by imprecise optical flow predictions as generation conditions, neglecting the influence of such misalignment on the preservation of appearance details. Moreover, most previous approaches employ generative adversarial networks (GANs) [17] for talking face generation, which often suffers from training instability and the issue of mode collapse.

To tackle the challenges above, we propose an innovative landmark-based diffusion model to learn the audio-visual relation. Figure 1 compares previous methods and ours. Our method utilizes facial landmarks as intermediate representation to ease the ambiguity of audio-visual mapping while enabling the end-to-end optimization of distinct stages, facilitating the generation of high-fidelity and lip-synced talking face video. Specifically, our two-stage approach initially converts the input audio signal into a set of landmark representations using a landmark completion module [11]. To enable the end-to-end optimization of distinct stages and align the synthesized motion with the motion represented by landmarks, we devise a novel conditioning module called TalkFormer to integrate landmark representations into the diffusion model using differentiable cross-attention. This end-to-end approach significantly reduces error accumulation resulting from pre-estimated landmark inaccuracies, thereby improving lip-audio synchronization. Besides, to enhance the fidelity of facial appearance, our approach converts a reference facial image into multi-scale features capturing intricate details. Then, our proposed TalkFormer module spatially aligns these features with the target motion using an implicit warping technique. Without using imprecise optical flow, the implicit warping automatically establishes semantic correlations for improved alignment between the reference features and the synthesized content. These aligned reference features are then integrated into the denoising process, enhancing the preservation of subject appearance details from reference image. Our extensive experiments validate our framework's effectiveness in producing realistic talking face videos with high-fidelity subject appearance and lip movements accurately synchronized to the input audio. We summarize our contributions as follows:

- We propose a novel method to learn the audio-visual relationship for generating high-fidelity and lip-synced talking face videos, utilizing facial landmarks as intermediate representation to ease the ambiguity of audio-visual mapping while enabling end-to-end optimization to minimize error accumulation.
- We introduce a novel conditioning module, TalkFormer, to align the synthesized motion with the motion represented by landmarks in a differentiable manner, enabling the joint optimization of distinct stages. Additionally, TalkFormer aligns the reference image features with the target motion based on semantic correlations, enhancing the preservation of subject appearance details.
- We conduct comprehensive experiments to demonstrate our method's effectiveness in producing high-fidelity and lip-synced talking face videos, which can generalize to any unseen subject without additional fine-tuning.

II. Related Work

A. Audio-Driven Talking Face Generation

Audio-driven talking face video generation techniques can be mainly divided into two types: person-specific or person-generic. Many person-specific methods can generate vivid videos [4], [5], [6], [7], [18], [19], but they require videos of the target subject for additional training. On the contrary, person-generic methods [1], [8], [9], [10], [11], [12], [13], [14], [20] enable inference on unseen subjects without any retraining or fine-tuning. However, there is still a gap in achieving high-fidelity and lip-synced talking face generation for person-generic methods.

Wav2Lip [8], PD-FGC [10], and PC-AVS [20] attempt to generate lip-synced videos by directly conditioning the generator on audio representation. Nevertheless, these approaches exhibit notable flaws and loss of appearance details in the subjects due to the inherent uncertainty and ambiguity in audio-visual mapping. IP-LAP [11] and other methods [1], [12] propose a two-stage framework that utilizes facial landmarks as intermediate representation. However, the projection of their intermediate landmark representation into the sketch image is indifferentiable. Subsequently, an image-to-image translation network is employed to synthesize realistic faces from the sketch images. Therefore, these methods train distinct stages independently and suffer from the inaccuracies of pre-estimated landmarks. Besides, prior methods [13], [14], [15], [21] including IP-LAP [11] utilize estimated optical flow to align the reference image with target facial expression and pose, such that more appearance details from reference image can be preserved during the generation process. However, accurate optical flow estimation is challenging especially when there is significant variation in head pose, leading to distorted

results. To tackle the drawbacks of GANs [17] in unstable training and mode collapse, DiffTalk [9] crafts a diffusion model to learn the direct audio-to-visual mapping for generalized talking face generation. It directly models the audio-to-lip translation with the landmarks of upper-half face concatenated as an auxiliary condition. However, the landmarks of upper-half face are insufficient to alleviate the uncertainty of direct audio-to-visual mapping. Besides, its usage of the misaligned reference image hinders the preservation of subject appearance details from reference image. Recently, GAIA [22] proposes to disentangle motion and appearance using VAE [23] and utilizes diffusion models to predict motion from the speech. However, it can not achieve end-to-end learning of the framework to reduce error accumulation. Besides, it leverages misaligned reference appearance features as generation conditions, which hinders the preservation of facial details from reference images.

Contrasting with prior approaches, our framework employs facial landmarks as intermediate representation while enabling end-to-end optimization. Our novel conditioning module TalkFormer integrates landmarks representation into diffusion models in a differentiable way and aligns the reference appearance features based on semantic correlations, facilitating the generation of high-fidelity and lip-synced talking face video.

B. Diffusion Models

Diffusion models [24], [25] have recently emerged as a promising type of generative models, exhibiting superior generation power and enhanced training stability when compared to GANs [17]. Denoising diffusion probabilistic model (DDPM) [25] is a class of latent variable models that uses a diffusion process to add noise to the data gradually and learns a denoiser network to reverse the diffusion process. Recently, Denoising Diffusion Implicit Models (DDIM) [26] have been proposed to accelerate the sampling process of diffusion models through a class of non-Markovian diffusion processes. Latent diffusion models (LDM) [27] apply diffusion model in the latent space of powerful pre-trained autoencoders to save computational resources while retaining the generation quality. Diffusion models have recently demonstrated remarkable success in various synthesis tasks [28], [29], [30], [31], [32], including text-to-image generation [27], [33], [34], image editing [35], [36], person image synthesis [37], [38], face video editing [28], and face restoration [30], [39]. Stable Diffusion [27] conditions the latent diffusion model on the CLIP [40] text embedding, achieving compelling text-to-image generation. ControlNet [33] proposes a neural network architecture that incorporates spatial conditions (e.g., sketch image) into pre-trained diffusion models.

For talking face video generation with diffusion models, previous methods [9], [41], [42], [43] make an early attempt to learn the direct audio-visual mapping by conditioning the diffusion model on audio feature, but often produce flawed results due to the ambiguity of mapping. To alleviate this problem, a straightforward method is to involve facial landmarks as intermediate representation and integrate the predicted landmarks from audio into the diffusion model using ControlNet [33] architecture. However, a notable drawback of this manner is the isolated training of audio-to-landmark and landmark-to-video stages, resulting in suboptimal lip synchronization. This is because the projection from landmarks coordinates into sketch image is indifferentiable. In contrast, in this study, we propose an innovative talking face video generation framework based on efficient latent diffusion models, which leverages facial landmarks as intermediate representation and enables end-to-end optimization of distinct stages. Our method first predicts the lip and jaw landmarks coordinates from audio signals, and a novel conditioning module TalkFormer integrates the landmarks into diffusion model in a differentiable manner.

III. Methodology

The framework overview of our method during inference and training is shown in Figure 2 and Figure 3, respectively. For talking face video generation, our method takes audio and a template video as input, masking the lower-half face. The framework then inpaints these areas with realistic content synchronized with the audio. Information about the subject appearance and facial contours is derived from a single reference image and reference full-face landmarks from the template video, respectively. More specifically, the audio signal drives the completion of lip and jaw landmarks, guided by reference full-face landmarks and pose landmarks detected from the upper-half face of the template video. The completed landmarks, along with the reference image, are then fed into the latent diffusion model via TalkFormer, influencing the synthesized motion and appearance. During training, the network diffuses the lower half of the ground-truth face in latent space, focusing on noise reduction. Upcoming sections Section III-A, Section III-B, and Section III-C will introduce latent diffusion models and provide a comprehensive explanation of our approach.

A. Preliminaries of Latent Diffusion Models

Latent diffusion models [27] carry out diffusion and the denoising process in the encoded latent space of an autoencoder $\mathcal{D}(\mathcal{E}(\cdot))$, with $\mathcal{E}(\cdot)$ being the encoder and $\mathcal{D}(\cdot)$ being the decoder. A U-Net-based [44] denoising network $\epsilon_\theta(z_t, t)$ is trained to predict the noise added to the image latent z_0, where $z_0 = \mathcal{E}(x)$, x is the input image, z_t represents the noisy version of z_0 at time step condition $t \in \{1, 2, \ldots, T\}$ and θ refers to the learnable parameters. The optimization objective during training is as follows:

$$\mathcal{L}_{ldm} = \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right] \tag{1}$$

where ϵ is the ground-truth noise added to the image latent z_0 and t is uniformly sampled from $\{1, \ldots, T\}$. During the inference phase, these models progressively denoise a normally distributed variable $z_T \sim \mathcal{N}(0, 1)$ until it reaches a clean latent $\hat{z}_0$. This clean latent variable can then be decoded by $\mathcal{D}$ to synthesize realistic images.

In our framework, both diffusion and denoising processes are exclusively performed in the lower half of the encoded latent, with the remaining upper half also being incorporated into denoising U-Net to provide more context. During inference, the masked input face is encoded as z_0^m, of which the

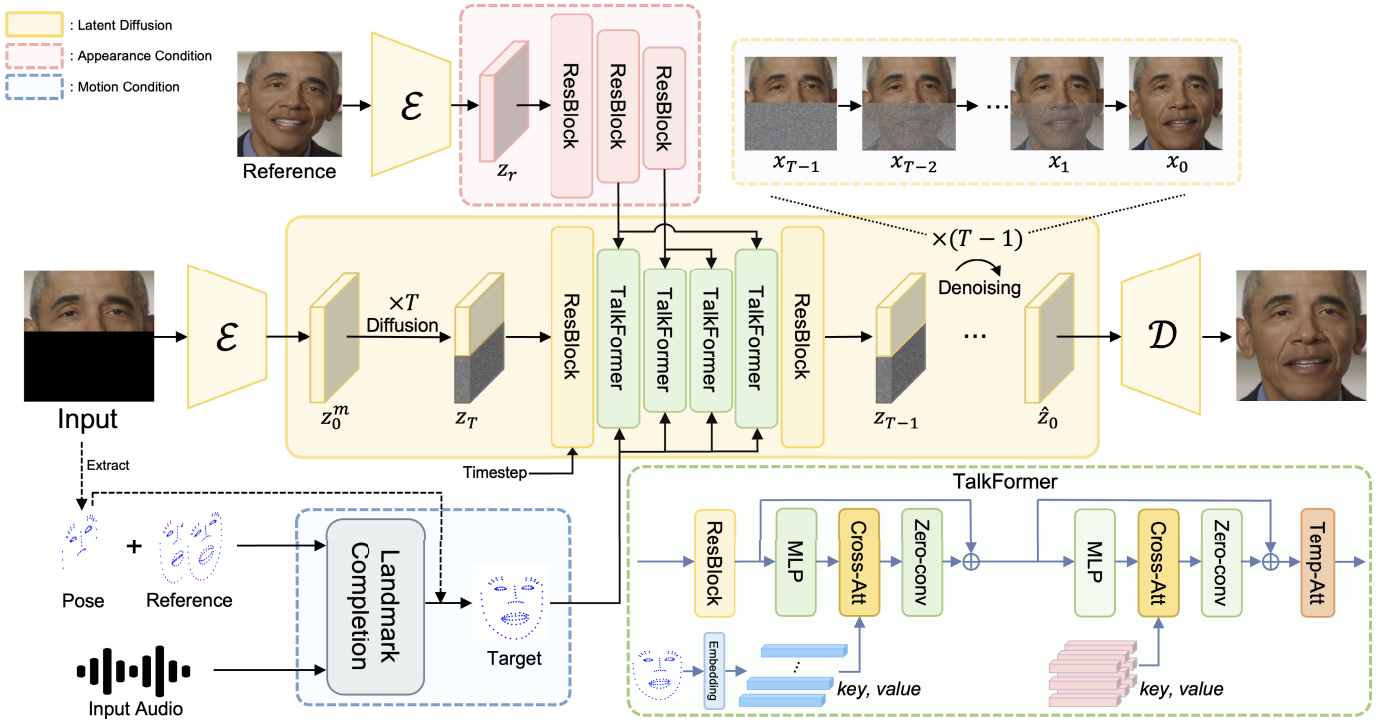

Fig. 2. An overview of the proposed framework during inference. The diffusion and reverse denoising operations are executed in the encoded latent space of an autoencoder $\mathcal{D}(\mathcal{E}(\cdot))$. (1). Initially, the audio signal drives the completion of lip and jaw landmarks, guided by reference full-face and upper half-face pose landmarks. The completed lip and jaw landmarks are then combined with the input pose landmarks to form the target full-face landmarks. (2). The conditioning module, TalkFormer, aligns the synthesized motion with the motion represented by target landmarks via differentiable cross-attention layers. To capture the intricate appearance details, a reference face image is encoded into multi-scale reference features. TalkFormer then aligns these features with the target motion via an implicit warping mechanism implemented by cross-attention layers. The skip-connections of U-Net are omitted for clarity.

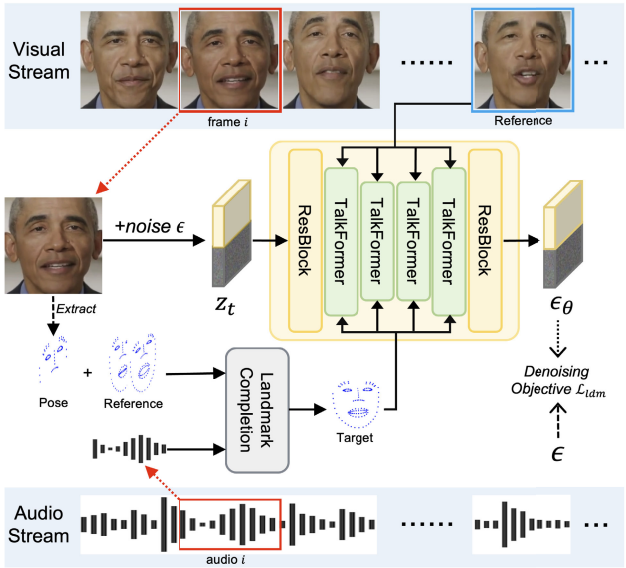

Fig. 3. The diagram of the training framework. Each training video contains a visual stream and an audio stream. During training, a frame i and its corresponding audio segment i are randomly selected from the video. The network learns to predict the noise added to the latent of the selected frame, guided by a reference image and landmarks sampled from different time steps in the same video.

lower half is diffused to obtain the initial z_T. During training, the ground-truth face is first encoded to z_0 and subsequently diffused to z_t by the noise.

B. Audio-Driven Landmark Completion

Instead of directly conditioning the diffusion model on the audio signal, our framework first establishes the less ambiguous mapping from audio to landmarks motion of lip and jaw. Following [11], we devise a transformer-based landmark completion module to predict the lip and jaw landmarks from the input audio. Specifically, we first encode the pose landmarks from the upper half of the face into a pose embedding using a 1D convolutional module. For referencing facial contour information, we extract reference full-face landmarks from N video frames within the input video, and encode them into N reference embeddings via another 1D convolutional module. The mel spectrogram of the input audio is encoded into an audio embedding by a 2D convolutional module. These pose, reference, and audio embeddings are concatenated along the sequence length dimension and subsequently fed into a transformer encoder to predict the lip and jaw landmark coordinates, denoted as $\hat{C}^{\tau}_{lip} \in \mathbb{R}^{2\times n_l}$ and $\hat{C}^{\tau}_{jaw} \in \mathbb{R}^{2\times n_j}$, respectively, where τ indicates landmarks for the τ-th frame, n_l and n_j are the number of landmarks used to represent the lip and jaw, respectively. To distinguish positional information and different modalities, sinusoidal positional embeddings and learnable modality embeddings are added to these embeddings. The transformer encoder comprises a series of multi-head self-attention layers, layer normalization, and MLP layers.

To ensure temporally stable landmark prediction, we adopt the batched sequential training strategy following the common practice of previous methods [45], [46], [47]. Specifically, the

completion module predicts landmarks of L successive frames for each video during training. The training objective for the landmark completion module is defined as follows:

$$\mathcal{L}_1 = \sum_{i=0}^{L-1} \left(\|\hat{C}_{lip}^{\tau+i} - C_{lip}^{\tau+i}\|_1 + \|\hat{C}_{jaw}^{\tau+i} - C_{jaw}^{\tau+i}\|_1 \right) \tag{2}$$

where $C_{lip}^{\tau+i}$ and $C_{jaw}^{\tau+i}$ are the ground-truth landmarks coordinates of lip and jaw, respectively. The predicted lip and jaw landmarks are then combined with the input pose landmarks to form the comprehensive target full-face landmarks.

However, the objective in Equation (2) is insufficient to synchronize the lip and jaw motion with input audio due to potential inaccuracies inherent in the pre-estimated ground-truth landmarks. Therefore, we expect the predicted landmarks to be integrated into the image generation stage (Section III-C) in a differentiable manner, enabling end-to-end optimization to improve lip synchronization.

C. Inpainting Lower Half via Latent Diffusion Model

As GANs [17] suffer from training instability and mode collapse, we resort to powerful latent diffusion models [27] to inpaint the lower half of the face, conditioning on the completed landmarks and reference image. For the end-to-end optimization of the whole framework and improved alignment between the reference image and the synthesized content, we introduce a novel conditioning module called TalkFormer, as illustrated in the green section of Figure 2. TalkFormer aligns the synthesized motion with the motion represented by facial landmarks via differentiable cross-attention [48], and aligns the reference image features via an implicit warping manner implemented by another cross-attention layer. In our denoiser U-Net, TalkFormer modules exist at all scales, except the first scale which only contains residual convolution blocks. In the following subsections, we will detail the core components of TalkFormer and the reference appearance encoder for encoding reference facial image.

1) TalkFormer: Align Talking Motion Differentiably: Previous researches [1], [11], [12] project the intermediate landmark representation on the image plane, forming the sketch image as a generation condition in an indifferentiable manner. In contrast, our TalkFormer first uses a 1D-convolution embedding module to encode the target full-face landmarks from the landmark completion module into n landmark embeddings $\{e_i, i = 1, 2, \ldots, n\}$, where n is the number of landmarks to represent the full face. Then, these landmark embeddings are integrated into cross-attention layers as keys and values, denoted as K_1 and V_1, respectively. Simultaneously, the queries Q_1 are extracted from the hidden features after ResNet [49] blocks through an MLP layer. The output of cross-attention is computed as Y according to the following equation:

$$Y = \text{Softmax}\left(\frac{Q_1 K_1^\top}{\sqrt{d_1}}\right) V_1 \tag{3}$$

where d_1 is the dimension of queries and keys. Subsequently, the results Y go through a zero-initialized convolution layer and are added to the hidden features of U-Net in a residual manner. In this way, the final generated face is ensured to have talking motion aligned with the motion represented by landmarks, and the diffusion model can be jointly optimized with the landmark completion module for improved lip synchronization.

2) Reference Appearance Encoder: To enable generalized talking face generation, a single reference face image is typically utilized as a condition, ensuring that the synthesized appearance remains consistent with the subject appearance. As illustrated in the pink section of Figure 2, the reference face image is initially encoded to the latent space as z_r. To retain more fine-grained details from the reference face image, we devise an appearance encoder similar to the U-Net encoder consisting of residual convolution blocks. This appearance encoder converts the latent z_r into multi-scale reference features symbolized as $F_a = \{F_a^i \mid i = 1, 2, .., I\}$, where I represents the number of scales in U-Net. The dimensions of these features are identical to those of the hidden features in the encoder of U-Net denoiser.

3) TalkFormer: Align Reference Appearance Features: To make the denoiser model aware of more appearance details from the reference image, we align the multi-scale reference appearance features in an implicit warping manner through another cross-attention layer. Specifically, we denote the hidden features after talking motion alignment as $F_h^i \in \mathbb{R}^{D \times H \times W}$, where scale $i \in \{2, 3, \ldots, I\}$. To spatially align the reference appearance features F_a^i with the hidden features F_h^i, the F_h^i is first transformed to the queries $Q_2 \in \mathbb{R}^{HW \times d_2}$ through an MLP layer while the F_a^i are projected to the keys $K_2 \in \mathbb{R}^{HW \times d_2}$ and values $V_2 \in \mathbb{R}^{HW \times D}$, where d_2 is the dimension of the keys. Then, the correlation matrix between F_a^i and F_h^i is computed as follows:

$$S = \text{Softmax}\left(\frac{Q_2 K_2^\top}{\sqrt{d_2}}\right) \tag{4}$$

where $S \in \mathbb{R}^{HW \times HW}$, and each element s_{jk} of it indicates the semantic correspondence between the hidden feature in location j and the reference feature in location k, with $j, k \in \{1, 2, \ldots, HW\}$. Based on this correlation matrix, we can obtain the aligned reference features by referring to the relevant features in the reference appearance features. Specifically, the reference appearance features F_a^i are warped implicitly via a weighted sum of the values in V_2 as follows:

$$\bar{F}_a^i = \text{Reshape}(S V_2) \tag{5}$$

where $\bar{F}_a^i \in \mathbb{R}^{D \times H \times W}$, and its semantic contents are spatially aligned with those of the hidden features F_h^i. Eventually, the aligned reference appearance features $\bar{F}_a^i$ are passed through a zero-initialized convolution layer and added to the hidden features F_h^i using a residual way. Consequently, the denoising process can better preserve the subject appearance details from reference images, facilitating high-fidelity talking face video generation.

Since our method focuses on talking face video generation, maintaining consistency across the generated video sequences over time is crucial. To this end, we follow Hallo [50] to incorporate a temporal attention layer into our TalkFormer module, as shown by the orange layer "Temp-Att" within TalkFormer in Figure 2. These layers perform self-attention along the temporal axis of the video frame sequence.

D. Optimization

Benefit from TalkFormer, the joint optimization of the landmark completion module and the latent diffusion model can be achieved by employing the following objective function:

$$\mathcal{L}_{total} = \mathcal{L}_{ldm} + \lambda \mathcal{L}_1 \tag{6}$$

where $\mathcal{L}_{ldm}$ is the denoising objective defined in Equation (1) and λ represents the weight assigned to the $\mathcal{L}_1$ loss term (Equation (2)). In this way, the $\mathcal{L}_{ldm}$ loss will guide the landmark completion module to predict more accurate landmarks for better denoising, thus enhancing lip-audio synchronization. Similar to the landmark completion module for improved temporal continuity, the latent diffusion model adopts the batched sequential training strategy [45], [46], [47] where L successive images are synthesized for each video during training.

IV. Experiments

A. Experimental Setups

1) Dataset: We conduct experiments on two public audio-visual datasets, VoxCeleb [51] and HDTF [14]. VoxCeleb is a collection of over 100,000 utterances from 1,251 celebrities, all extracted from videos uploaded to YouTube. HDTF is a high-resolution audio-visual dataset consisting of approximately 362 distinct videos, spanning over 15.8 hours, in 720P or 1080P resolutions. Compared to HDTF, the large-scale VoxCeleb dataset is a more standard benchmark commonly used in prior work. To ensure a fair comparison, all comparison methods, including ours, are trained on the VoxCeleb dataset and evaluated using the test sets of both VoxCeleb and HDTF.

2) Evaluation Metric: We quantitatively evaluate all methods regarding visual quality and lip synchronization. Pixel-level visual quality is assessed through the Peak Signal-to-Noise Ratio (PSNR) and Structured Similarity (SSIM) [52], while feature-level visual quality is evaluated using Learned Perceptual Image Patch Similarity (LPIPS) [53] and Fréchet Inception Distance (FID) [54]. Compared to pixel-level measurements, the feature-level measurements are more in line with human perception [53], [55]. Additionally, we employ the cosine similarity (CSIM) of identity vectors extracted by the ArcFace face recognition network [56] to assess the preservation of subject identity. SyncScore [57] is commonly used by prior work to evaluate the lip-audio synchronization quality, despite some limitations.

3) Comparison Methods: We compare our approach against several state-of-the-art person-generic audio-driven talking face video generation methods. DiffTalk (*CVPR'23*) [9] constructs a Diffusion-based framework for generalized talking face synthesis by conditioning the latent diffusion model on audio signal. PD-FGC (*CVPR'23*) [10] employs a progressive disentangled representation learning strategy to achieve fine-grained controllable talking face synthesis (e.g., eye, pose control). IP-LAP (*CVPR'23*) [11] is a two-stage landmark-based method that trains different stages separately and utilizes predicted optical flow to align the reference image with the target pose and expression. PC-AVS (*CVPR'21*) [20] proposes a GAN-based framework to generate pose-controllable talking face videos by modularizing audio-visual representations. Wav2Lip (*MM'20*) [8] utilizes a lip sync discriminator to guide the generator in generating lip-synced talking face videos.

4) Comparison Setups: Wav2Lip [8], IP-LAP [11], DiffTalk [9], and our method all generate talking face videos by inpainting the lower half of the face. Therefore, during the quantitative comparison, the lower half of the face in the input video is masked. Then, these methods reconstruct the masked area guided by the input audio and reference image. The original input video serves as the ground truth for metric calculation. We train DiffTalk [9] using the official code until convergence, but it generates temporally unstable results. It relies on additional frame interpolation to smooth the results, affecting comparison fairness. Therefore, the frame interpolation post-processing was not employed for fair comparison. PC-AVS [20] utilizes a pose source video, an audio input, and a reference image to generate a talking face video. In our implementation, we substitute its pose source video with the ground-truth video. PD-FGC [10] requires a pose source video, an expression source video, an eye blink source video, an audio input, and a reference image to generate a talking face video. Our implementation replaces its pose source, expression source, and eye blink source with the ground-truth video.

5) Implementation Details: In our framework, the input face images are first cropped from the video and then resized to 256×256 resolution, and the latent space of autoencoder $\mathcal{D}(\mathcal{E}(\cdot))$ has a spatial dimension of 64×64. Facial landmarks are extracted from video frames using the mediapipe tool [58]. We represent the lip with $n_l = 41$ landmarks, the jaw with $n_j = 16$ landmarks, and the entire face with a total of $n = 131$ landmarks. All landmarks are arranged as arrays in a predefined order. We set N to 5 and the reference full-face landmarks are detected from the randomly selected frames of input videos. The hyper-parameter λ is set to 10. Following previous works [8], [11], L is set to 5. To generate talking face videos, we employ the DDIM [26] diffusion sampler with 200 steps. We set the number of diffusion steps T to 1000. The number of scales I is 4, but we illustrate the case of $I = 3$ in Figure 2 for clarity. The reference face image can be any face image from the input video that reflects as many appearance details of the subject as possible. All comparison methods use the same reference face image to ensure a fair evaluation. Our implementation of the proposed method closely follows the code implementation of latent-diffusion [59], while incorporating TalkFormer, Appearance Encoder, and Landmark Completion module [11] as additional components. The Appearance Encoder is designed based on the encoder structure of U-Net denoiser in latent-diffusion [59], excluding self-attention layers. The temporal attention layer within our TalkFormer module follows the implementation of Hallo [50], where two motion frames from the preceding time are encoded by the Appearance Encoder and then concatenated with the hidden features for temporal self-attention.

We train our framework on two NVIDIA A100 (40GB) GPUs for 500 epochs with Adam optimizer [60]. The batch size is 64, and the learning rate is 4e-5. The pre-trained

TABLE I

QUANTITATIVE COMPARISONS BETWEEN THE PROPOSED AND PREVIOUS STATE-OF-THE-ART METHODS IN PERSON-GENERIC TALKING FACE VIDEO GENERATION.HERE ↑ DENOTES HIGHER IS BETTER, AND ↓ INDICATES LOWER IS BETTER

Method	VoxCeleb						HDTF					
	PSNR↑	SSIM↑	LPIPS↓	FID↓	CSIM↑	SyncScore↑	PSNR↑	SSIM↑	LPIPS↓	FID↓	CSIM↑	SyncScore↑
Ground Truth	-	-	-	-	-	8.59	-	-	-	-	-	8.87
Wav2Lip [8]	24.17	0.81	0.173	78.20	0.52	**9.33**	22.51	0.78	0.232	87.99	0.60	**9.43**
PC-AVS [20]	21.84	0.76	0.132	53.72	0.37	8.31	19.18	0.68	0.189	56.33	0.33	8.73
IP-LAP [11]	24.17	0.83	0.114	44.02	0.55	3.39	22.60	0.83	0.118	34.92	0.67	3.62
PD-FGC [10]	20.15	0.70	0.165	57.54	0.33	6.33	17.76	0.65	0.207	62.54	0.31	6.47
DiffTalk [9]	22.43	0.74	0.119	44.00	0.51	1.42	22.03	0.74	0.121	30.16	0.56	2.30
Ours	**25.44**	**0.85**	**0.090**	**37.19**	**0.57**	4.42	**23.48**	**0.83**	**0.096**	**27.94**	**0.69**	5.03

autoencoder is frozen during training. The Landmark Completion module is jointly trained from scratch with the latent diffusion model. Our method focuses on inpainting the lower half of the face based on the input audio. Hence, to generate talking face videos, we first crop the face area from the input template video as network input. After obtaining the generated face, we employ a post-processing technique following the previous method [11] to seamlessly blend it with the background via a Gaussian-smoothed face mask, producing final talking face videos. For fair comparison, this post-processing technique was not employed during the quantitative and qualitative comparisons. The input template video has no length requirement and can be looped to match the length of the input audio.

B. Quantitative Evaluation

We conduct a comprehensive quantitative comparison with state-of-the-art methods regarding visual quality and lip synchronization. The visual quality metrics are calculated solely based on the lower half of the generated face, since the generated upper-half face in Wav2Lip [8], IP-LAP [11], DiffTalk [9], and ours almost inherit from the input video (i.e., ground truth). The comparison results are reported in Table I.

1) Visual Quality: Our method outperforms other methods in all visual quality metrics (PSNR, SSIM, LPIPS, FID, CSIM) on both VoxCeleb [51] and HDTF [14] datasets. Specifically, on the perceptual distance metrics LPIPS and FID, our method significantly improves over other methods. This verifies that our method can produce high-fidelity talking face videos that align with human perception, preserving more appearance details. Besides, the highest CSIM score achieved by ours also indicates our method can preserve more identity information of the target subject. Although Wav2Lip exhibits a slight lag in terms of PSNR and SSIM metrics compared to our method, its FID and LPIPS values are approximately twice as high as ours, suggesting the presence of artifacts that are not aligned with human perception in their results. The performance of IP-LAP is closely comparable to ours in terms of the PSNR, SSIM, and CSIM metrics, but there remains a certain gap between IP-LAP and ours when assessed on the LPIPS, FID, and SyncScore metrics. While DiffTalk exhibits comparable performance in the FID metric, it still lags behind our approach when assessed on the LPIPS and CSIM metrics.

2) Lip Synchronization: Due to different speaking styles among individuals, accurately assessing lip-audio synchronization remains a persistently challenging task. A common practice is to calculate the SyncScore based on the audio and visual features of SyncNet [57]. Wav2Lip, PC-AVS, and PD-FGC directly model the audio-visual mapping and obtain a better SyncScore than ours. However, our approach notably excels in visual quality metrics, particularly in preserving the finer details of subject appearance, an aspect where others have room for improvement. Wav2Lip utilizes SyncNet [57] as a discriminator during training. Hence, it achieves a very high SyncScore, even higher than that of the ground truth. Besides, PD-FGC and PC-AVS adopt audio-visual contrastive learning similar to SyncNet [57], which contributes to a higher SyncScore but compromises visual quality. IP-LAP leverages facial landmarks as intermediate representations, but its lip synchronization is inferior to ours due to the isolated training of different stages. DiffTalk directly models the audio-visual mapping and generates temporally unstable videos with poor lip synchronization. We suspect its issue might stem from the mapping ambiguity magnified by the multi-step iteration of diffusion model.

C. Qualitative Evaluation

1) Visual Comparison: As shown in Figure 4, we present some representative comparison results on the HDTF [14] and VoxCeleb [51] datasets. It can be observed that our results are visually closer to the ground-truth images than other methods', with more appearance details (e.g., beard, lip, teeth) preserved. It implies that our TalkFormer module could effectively align the reference image features based on semantic correlation, providing valuable features for the diffusion model. Besides, the lip shapes of ours are also closer to the ground truth. DiffTalk, PD-FGC, PC-AVS, and Wav2Lip directly learn the audio-visual mapping and utilize a misaligned reference image as a generation condition. Therefore, their results lose some appearance details of the subject and appear blurry. IP-LAP generates results that are blurrier than ours and exhibit some artifacts, possibly due to the inaccurate optical flow estimation for aligning reference images. For more qualitative comparisons, please refer to the supplementary video as detailed in the following subsection.

Fig. 4. Several representative visual comparisons. The subject on the left is from the VoxCeleb [51] dataset, while the subject on the right is from the HDTF [14] dataset. Our method achieves high fidelity of the subject appearance details with accurate lip shape. For more qualitative results, please refer to the supplementary video.

2) Supplementary Video: We have provided a short video as supplementary material in this link. The time schedule of the video is as follows:

- 00:00 ~ 00:08: Video title.
- 00:08 ~ 00:46: Demonstration of Ours. We demonstrate a generated result of our method where the driving audio is sourced from the text-to-speech technique, and the subject is from the HDTF [14] dataset. We also visualize the intermediate landmarks after the landmark completion module.
- 00:46 ~ 01:42: Method Comparison. We present the results of all methods as well as ground-truth videos for comparison.
- 01:42 ~ 02:19: Ablation Study. We present the qualitative results of the ablation study, which will be further analyzed in the Section IV-D.
- 02:19 ~ 03:14: More results of our method. We provide the testing results of our method for more cases.

3) User Study: For comprehensive evaluation, we conduct a user study where 16 volunteers are invited to assess the generated videos of all comparison methods. We randomly sample 10 videos for testing, five from the HDTF [14] dataset and five from the VoxCeleb [51] dataset. Volunteers are asked to give their ratings (0-(5) for each generated video regarding image quality, lip-audio synchronization, and fidelity of appearance details. The videos are presented to participants in a random order and the evaluation criteria are explained in detail to them. The mean opinion scores (MOS) of each method are presented in Table II. Our method receives better evaluations from participants across three dimensions than other approaches.

TABLE II
USER STUDY RESULTS MEASURED BY MEAN OPINION SCORES. SCORES RANGE FROM 0 TO 5, WHERE HIGHER SCORES DENOTE SUPERIOR PERFORMANCE

Method	Appearance Fidelity	Lip Synchronization	Image Quality
Wav2Lip [8]	2.85	3.67	2.82
PC-AVS [20]	2.97	3.17	2.93
IP-LAP [11]	2.88	2.88	2.79
PD-FGC [10]	2.76	3.19	2.76
DiffTalk [9]	2.52	1.24	2.01
Ours	**4.55**	**4.32**	**4.54**

D. Ablation Study

In this section, we conduct an ablation study on the HDTF [14] dataset to verify the effectiveness of the proposed end-to-end framework and TalkFormer conditioning module. The numerical results are reported in Table III, while the qualitative

Fig. 5. Ablation study on the effectiveness of end-to-end optimization and the TalkFormer module. "Ours w/o End2End" represents the variant without end-to-end training. "Ours w/o M-Align" represents the variant without talking motion alignment in TalkFormer. "Ours w/o R-Align" represents the variant without reference appearance features alignment in TalkFormer.

TABLE III

ABLATION STUDY RESULTS OF REMOVING INDIVIDUAL COMPONENTS OF THE PROPOSED APPROACH.THE TERMS "OURS W/O M-ALIGN", "OURS W/O R-ALIGN", "OURS W/O B-SEQUENTIAL", AND "OURS W/O END2END" DENOTE VARIANTS OF OUR MODEL, INDICATING THE ABSENCE OF TALKING MOTION ALIGNMENT, REFERENCE APPEARANCE FEATURE ALIGNMENT, BATCHED SEQUENTIAL TRAINING STRATEGY, AND END-TO-END TRAINING, RESPECTIVELY

Variants	PSNR↑	SSIM↑	LPIPS↓	FID↓	CSIM↑	SyncScore↑
Ours w/o M-Align	20.82	0.75	0.159	50.28	0.687	1.12
Ours w/o R-Align	24.28	0.82	0.114	43.36	0.790	1.54
Ours w/o B-Sequential	23.88	0.81	0.118	35.91	0.777	5.35
Ours w/o End2End	23.85	0.82	0.096	30.96	0.786	4.93
Ours	24.49	0.83	0.097	29.15	**0.795**	5.46
Ours + CompleteRef	25.12	0.86	0.091	23.66	0.794	5.48
Ours + TempAttn	**26.34**	**0.86**	**0.083**	**23.62**	0.791	**5.59**

results are presented in Figure 5, as well as in the supplementary video.

1) Effect of End-to-End Training: Our two-stage landmark-based method achieves the joint optimization of landmark completion module and latent diffusion model to improve lip synchronization. To validate the effectiveness of end-to-end optimization, we devise a variant where these two modules are trained separately. Specifically, the landmark completion module is optimized using only $\mathcal{L}_1$ loss (Equation (2)). In the denoiser U-Net of latent diffusion model, TalkFormer accepts the pre-estimated ground-truth landmarks as condition. The latent diffusion model is then optimized using only $\mathcal{L}_{ldm}$ loss (Equation (1)).

The numerical results of this variant are reported in the "Ours w/o End2End" row of Table III. In terms of visual quality metrics, this variant exhibits similar performance to our full model. However, the SyncScore of this variant drops 9.71% compared to the full model's, verifying the effectiveness of end-to-end training in improving lip-audio synchronization. Besides, as seen in the "Ours w/o End2End" row of Figure 5, without the end-to-end optimization, the synthesized lip shapes are less accurate, while the visual quality remains similar to the full model's.

2) Effect of Talking Motion Alignment in TalkFormer: Our TalkFormer module aligns the synthesized motion with the motion represented by landmarks via a cross-attention layer. To implement a variant without talking motion alignment, we remove the first cross-attention layer in TalkFormer that integrates landmarks embeddings as condition. Following the practice of DiffTalk [9], we redesign the landmark embedding module composed of multiple MLP layers to encode the target full-face landmarks into a single landmark embedding. This landmark embedding is added to all spatial locations of the hidden features in U-Net.

The numerical results of this variant are reported in the "Ours w/o M-Align" row of Table III. It can be seen that the SyncScore drops significantly compared to the full model's. This is because the synthesized motion of the lip and jaw could not be accurately controlled through a simple addition operation. Besides, the addition operation may introduce artifacts into the generated results, deteriorating the visual quality metrics. As shown in the "Ours w/o M-Align" row of Figure 5, in the absence of TalkFormer's talking motion alignment, the mouth shape remains predominantly closed. Besides, influenced by the simple addition operation, the generated mouths appear blurry and exhibit some artifacts. In the first and fourth images of the "Ours w/o M-Align" row, the generated mouth shapes are somewhat skewed, which implies loss of the subject identity information.

3) Effect of Reference Features Alignment in TalkFormer: To develop a variant without reference appearance features alignment, we remove the second cross-attention layer in TalkFormer that incorporates the reference appearance features, and replace it with a self-attention layer akin to DDPM [25]. Besides, the appearance encoder is removed. Following the common practice of previous researches [8], [9], the reference face image is first encoded into the latent space as z_r. The z_r is then concatenated with the noisy latent z_t along the channel dimension and fed into the U-Net denoiser network.

The numerical results of this variant are reported in the "Ours w/o R-Align" row of Table III. It can be seen that all the visual quality metrics deteriorate compared to the full model's. Although the pixel-level metrics (PSNR, SSIM) do not change significantly, the feature-level metrics (LPIPS, FID), which are more in line with human perception, increase by a large margin. The potential reason is that the diffusion model can not extract meaningful features from the misaligned reference features, resulting in the loss of subject appearance details. Besides, the CSIM metric based on identity vectors might not be sensitive to subject appearance details, therefore the CSIM decreases slightly without reference features alignment in TalkFormer. Moreover, the lip shape of the misaligned reference image might have a negative impact on the synthesized lip shape. Therefore, the SyncScore decreases without reference features alignment. Furthermore, as can be seen in

TABLE IV
ABLATION STUDY ON THE NUMBER OF REFERENCE FULL-FACE LANDMARKS N

Variants	PSNR↑	SSIM↑	LPIPS↓	FID↓	CSIM↑	SyncScore↑
Ours (N=1)	23.66	0.80	0.111	31.74	0.781	4.55
Ours (N=3)	24.33	0.82	0.098	29.83	0.793	5.25
Ours (N=5)	**24.49**	0.83	**0.097**	**29.15**	**0.795**	5.46
Ours (N=10)	24.31	**0.84**	0.098	29.35	0.794	**5.48**

the "Ours w/o R-Align" row of Figure 5, in the absence of reference appearance features alignment, the generated results lose some subject appearance details, resulting in some unrealistic contents. Besides, the lip shapes are less accurate influenced by the misaligned reference face image.

4) Effect of Temporal Modeling Mechanism: Our method incorporates two temporal modeling mechanisms: the batched sequential training strategy and the temporal attention layer within the TalkFormer module. To evaluate the effectiveness of the batched sequential training strategy, we design a model variant without it. As shown in the "Ours w/o B-Sequential" row in Table III, this variant performs worse than our full model ("Ours") across all metrics, with particularly notable degradation in LPIPS and FID scores. To assess the effectiveness of the additional temporal attention layer, we further integrate it into the TalkFormer module of our method. The corresponding results are reported in the "Ours + TempAttn" row in Table III. This integration improves performance on all visual quality and lip-sync metrics except CSIM which remains largely unchanged. This demonstrates the benefit of temporal attention for enhancing temporal coherence and generation quality.

5) Effect of Reference Full-Face Landmarks: The landmark completion module extracts reference full-face landmarks from N video frames within the input video. To investigate the effect of different values of N, we conduct experiments by varying N, and the corresponding results are shown in Table IV. As can be observed, the model performance consistently improves as N increases. However, the performance gain becomes marginal when $N > 5$, indicating that reference landmarks from 5 frames are already sufficient to provide the necessary contextual information. Therefore, we set $N = 5$ in our experiments.

6) Effect of Design of Reference Appearance Encoder: The Reference Appearance Encoder in our method uses only the encoder part of a U-Net, while some recent works [50], [61] employ the complete U-Net structure to encode the reference image. To explore the impact of different architectural choices for the Reference Appearance Encoder, we conduct an additional ablation study using the complete U-Net structure. The corresponding results are shown in the "Ours + CompleteRef" row in Table III. The complete U-Net brings slight performance improvements, but it introduces additional parameters and computational overhead. Therefore, we recommend using the complete U-Net structure as the Reference Appearance Encoder when computational resources permit, while the encoder-only version is preferable in resource-constrained scenarios.

V. CONCLUSION AND DISCUSSION

In this paper, we propose a novel landmark-based framework to learn the audio-visual relationship for person-generic talking face video generation. Our framework utilizes facial landmarks as intermediate representations to alleviate the ambiguity of audio-visual mapping, while enabling end-to-end optimization to minimize error accumulation resulting from the inaccuracies of pre-estimated facial landmarks. This accomplishment can be attributed to our innovative conditioning module, TalkFormer, which aligns the synthesized talking motion with the motion represented by landmarks using a differentiable cross-attention layer. Besides, TalkFormer implicitly aligns the reference appearance features with the target motion based on semantic correlations, facilitating the preservation of more appearance details for the target subject. Extensive experiments have verified the effectiveness of our method in producing high-fidelity and lip-synced talking face videos.

ETHICAL DISCUSSION

Our method can generate a talking face video for any subject, requiring only a template video of the subject and a segment of audio, without the need for person-specific training. It may be misused for illicit gains. To combat the malicious behaviors, we will watermark the generated results and are willing to contribute our synthetic videos to the deepfake detection community for enhancing their algorithms.

REFERENCES

[1] T. Xie et al., "Towards realistic visual dubbing with heterogeneous sources," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1739–1747.
[2] Z. Zhou et al., "DialogueNeRF: Towards realistic avatar face-to-face conversation video generation," 2022, *arXiv:2203.07931*.
[3] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, "Neural voice puppetry: Audio-driven facial reenactment," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 716–731.
[4] Y. Guo, K. Chen, S. Liang, Y.-J. Liu, H. Bao, and J. Zhang, "AD-NeRF: Audio driven neural radiance fields for talking head synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5784–5794.
[5] R. Huang, P. Lai, Y. Qin, and G. Li, "Parametric implicit face representation for audio-driven facial reenactment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 12759–12768.
[6] C. Du et al., "DAE-Talker: High fidelity speech-driven talking face generation with diffusion autoencoder," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 4281–4289.
[7] Z. Ye, Z. K. Jiang, Y. Ren, J. Liu, J. He, and Z. Zhao, "GeneFace: Generalized and high-fidelity audio-driven 3D talking face synthesis," in *Proc. 11th Int. Conf. Learn. Represent.*, 2023.
[8] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 484–492.
[9] S. Shen et al., "DiffTalk: Crafting diffusion models for generalized audio-driven portraits animation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1982–1991.
[10] D. Wang, Y. Deng, Z. Yin, H.-Y. Shum, and B. Wang, "Progressive disentangled representation learning for fine-grained controllable talking head synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 17979–17989.
[11] W. Zhong et al., "Identity-preserving talking face generation with landmark and appearance priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9729–9738.
[12] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "MakeItTalk: Speaker-aware talking-head animation," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 1–15, Dec. 2020.

[13] S. Gururani, A. Mallya, T.-C. Wang, R. Valle, and M.-Y. Liu, "SPACE: Speech-driven portrait animation with controllable expression," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 20857–20866.

[14] Z. Zhang, L. Li, Y. Ding, and C. Fan, "Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3660–3669.

[15] S. Sinha, S. Biswas, R. Yadav, and B. Bhowmick, "Emotion-controllable generalized talking face generation," 2022, *arXiv:2205.01155*.

[16] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. 26th Annu. Conf. Comput. Graph. Interact. Techn. - SIGGRAPH*, 1999, pp. 187–194.

[17] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[18] S. Shen, W. Li, Z. Zhu, Y. Duan, J. Zhou, and J. Lu, "Learning dynamic facial radiance fields for few-shot talking head synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 666–682.

[19] C. Zhang et al., "FACIAL: Synthesizing dynamic talking face with implicit attribute learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3867–3876.

[20] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu, "Pose-controllable talking face generation by implicitly modularized audio-visual representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4176–4186.

[21] W. Zhang et al., "SadTalker: Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 8652–8661.

[22] T. He et al., "GAIA: Data-driven zero-shot talking avatar generation," in *Proc. 12th Int. Conf. Learn. Represent.*, 2024.

[23] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[24] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.

[25] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NIPS*, vol. 33, 2020, pp. 6840–6851.

[26] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Represent.*, 2020.

[27] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 10684–10695.

[28] G. Kim, H. Shim, H. Kim, Y. Choi, J. Kim, and E. Yang, "Diffusion video autoencoders: Toward temporally consistent face video editing via disentangled video encoding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 6091–6100.

[29] X. Xing, C. Wang, H. Zhou, J. Zhang, Q. Yu, and D. Xu, "DiffSketcher: Text guided vector sketch synthesis through latent diffusion models," 2023, *arXiv:2306.14685*.

[30] Y. Zhao, T. Hou, Y.-C. Su, X. Jia, Y. Li, and M. Grundmann, "Towards authentic face restoration with iterative diffusion models and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 7278–7288.

[31] J. Z. Wu et al., "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 7623–7633.

[32] J. Seo, G. Lee, S. Cho, J. Lee, and S. Kim, "MIDMs: Matching interleaved diffusion models for exemplar-based image translation," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2191–2199.

[33] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 3836–3847.

[34] C. Mou et al., "T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," 2023, *arXiv:2302.08453*.

[35] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-or, "Prompt-to-prompt image editing with cross-attention control," in *Proc. 11th Int. Conf. Learn. Represent.*, 2022.

[36] B. Yang et al., "Paint by example: Exemplar-based image editing with diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18381–18391.

[37] A. K. Bhunia et al., "Person image synthesis via denoising diffusion model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 5968–5976.

[38] F. Shen, H. Ye, J. Zhang, C. Wang, X. Han, and W. Yang, "Advancing pose-guided image synthesis with progressive conditional diffusion models," 2023, *arXiv:2310.06313*.

[39] Z. Yue and C. C. Loy, "DifFace: Blind face restoration with diffused error contraction," 2022, *arXiv:2212.06512*.

[40] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[41] M. Stypułkowski, K. Vougioukas, S. He, M. Zięba, S. Petridis, and M. Pantic, "Diffused heads: Diffusion models beat GANs on talking-face generation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 5089–5098.

[42] S. Mukhopadhyay, S. Suri, R. T. Gadde, and A. Shrivastava, "Diff2Lip: Audio conditioned diffusion models for lip-synchronization," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 5280–5290.

[43] D. Bigioi et al., "Speech driven video editing via an audio-conditioned diffusion model," *Image Vis. Comput.*, vol. 142, Feb. 2024, Art. no. 104911.

[44] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[45] S. Wang, L. Li, Y. Ding, and X. Yu, "One-shot talking face generation from single-speaker audio-visual correlation learning," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 3, pp. 2531–2539.

[46] Y. Ma et al., "StyleTalk: One-shot talking head generation with controllable speaking styles," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 1896–1904.

[47] S. Wang et al., "StyleTalk++: A unified framework for controlling the speaking styles of talking heads," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 6, pp. 4331–4347, Jun. 2024.

[48] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[50] M. Xu et al., "Hallo: Hierarchical audio-driven visual synthesis for portrait image animation," 2024, *arXiv:2406.08801*.

[51] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, Aug. 2017, pp. 2616–2620.

[52] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[53] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[54] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6626–6637.

[55] R. Zhen, W. Song, Q. He, J. Cao, L. Shi, and J. Luo, "Human–computer interaction system: A survey of talking-head generation," *Electronics*, vol. 12, no. 1, p. 218, Jan. 2023.

[56] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[57] J. S. Chung and A. Zisserman, "Out of time: Automated lip sync in the wild," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, Nov. 2016, pp. 251–263.

[58] C. Lugaresi et al., "MediaPipe: A framework for building perception pipelines," 2019, *arXiv:1906.08172*.

[59] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. (2022). *Latent-Diffusion*. [Online]. Available: https://github.com/CompVis/latent-diffusion

[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[61] L. Tian, Q. Wang, B. Zhang, and L. Bo, "Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2024, pp. 244–260.

Weizhi Zhong received the B.E. and M.E. degrees from the School of Computer Science and Engineering, Sun Yat-sen University, in 2022 and 2025, respectively. He is currently pursuing the Ph.D. degree with The University of Hong Kong. His research interests include computer vision and deep learning.

Junfan Lin received the Ph.D. degree in software engineering from Sun Yat-sen University, Guangzhou, China. He is currently a Postdoctoral Researcher with the Peng Cheng Laboratory, advised by Prof. Liang Lin. His research interests include reinforcement learning, embodied artificial intelligence, and causal inference.

Peixin Chen received the B.E. degree from the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China, in 2024, where he is currently pursuing the M.S. degree in computer science. His research interests include computer vision and machine learning.

Feng Gao (Member, IEEE) received the B.S. degree in computer science from University College London in 2007 and the Ph.D. degree in computer science from Peking University in 2018. He was a Postdoctoral Research Fellow with the Future Laboratory, Tsinghua University, from 2018 to 2020. He has been with Peking University as an Assistant Professor, since 2020. His research interests include intersection of computer science and art, including but not limited on artificial intelligence and painting art, deep learning, and painting robot.

Liang Lin (Fellow, IEEE) is currently a Full Professor of computer science with Sun Yat-sen University. He served as the Executive Director and the Distinguished Scientist of SenseTime Group from 2016 to 2018, leading the Research and Development teams for cutting-edge technology transferring. He has authored or co-authored more than 300 papers in leading academic journals and conferences, and his papers have been cited by more than 34 000 times. He is a fellow of IAPR. He was a recipient of numerous awards and honors including the Wu Wen-Jun Artificial Intelligence Award, the First Prize of China Society of Image and Graphics, the ICCV Best Paper Nomination in 2019, the Annual Best Paper Award by *Pattern Recognition* (Elsevier) in 2018, the Best Paper Diamond Award in IEEE ICME in 2017, and the Google Faculty Award in 2012. His supervised Ph.D. students received the ACM China Doctoral Dissertation Award, the CCF Best Doctoral Dissertation, and the CAAI Best Doctoral Dissertation. He served as the Area Chairs for numerous conferences, such as CVPR, ICCV, SIGKDD, and AAAI. He is an Associate Editor for IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS and IEEE TRANSACTIONS ON MULTIMEDIA.

Guanbin Li (Member, IEEE) received the Ph.D. degree from The University of Hong Kong in 2016. He is currently a Full Professor with the School of Data and Computer Science, Sun Yat-sen University. He has authorized and co-authorized on more than 200 papers in top-tier academic journals and conferences. His current research interests include computer vision, image processing, and deep learning. He was a recipient of ICCV 2019 Best Paper Nomination Award.