StableFace: Analyzing and Improving Motion Stability for Talking Face Generation

Jun Ling, Xu Tan, Senior Member, IEEE, Liyang Chen, Runnan Li, Yuchao Zhang, Sheng Zhao, Senior Member, IEEE, and Li Song, Senior Member, IEEE

Abstract—While previous methods for speech-driven talking face generation have shown significant advances in improving the visual and lip-sync quality of the synthesized videos, they have paid less attention to lip motion jitters which can substantially undermine the perceived quality of talking face videos. What causes motion jitters, and how to mitigate the problem? In this article, we conduct systematic analyses to investigate the motion jittering problem based on a state-of-the-art pipeline that utilizes 3D face representations to bridge the input audio and output video, and implement several effective designs to improve motion stability. This study finds that several factors can lead to jitters in the synthesized talking face video, including jitters from the input face representations, training-inference mismatch, and a lack of dependency modeling in the generation network. Accordingly, we propose three effective solutions: 1) a Gaussian-based adaptive smoothing module to smooth the 3D face representations to eliminate jitters in the input; 2) augmented erosions added to the input data of the neural renderer in training to simulate the inference distortion to reduce mismatch; 3) an audio-fused transformer generator to model inter-frame dependency. In addition, considering there is no off-the-shelf metric that can measure motion jitters of talking face video, we devise an objective metric (Motion Stability Index, MSI) to quantitatively measure the motion jitters. Extensive experimental results show the superiority of the proposed method on motion-stable talking face generation, with superior quality to previous systems.

Index Terms—Talking face generation, vision transformer, motion jitters, motion stability index.

Manuscript received 29 August 2022; revised 5 April 2023 and 9 October 2023; accepted 6 November 2023. Date of publication 16 November 2023. Date of current version 28 December 2023. This work was supported by the Fundamental Research Funds for the Central Universities, STCSM under Grant 22DZ2229005, 111 Project BP0719010. The work of Jun Ling was conducted at Microsoft. The guest editor coordinating the review of this manuscript and approving it for publication was Mr. Sebastian Bosse. (Corresponding author: Li Song.)

Jun Ling is with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: lingjun@sjtu.edu.cn).
Xu Tan is with Microsoft Research Asia (MSRA), Beijing 100080, China (e-mail: xuta@microsoft.com).
Liyang Chen is with Microsoft Research Asia (MSRA), Beijing 100080, China (e-mail: cly21@mails.tsinghua.edu.cn).
Runnan Li, Yuchao Zhang, and Sheng Zhao are with the Microsoft Cloud+AI, Beijing 100080, China (e-mail: runnan.li@microsoft.com; yuchao.zhang@microsoft.com; sheng.zhao@microsoft.com).
Li Song is with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: song_li@sjtu.edu.cn).
This article has supplementary downloadable material available at https://doi.org/10.1109/JSTSP.2023.3333552, provided by the authors.
Digital Object Identifier 10.1109/JSTSP.2023.3333552

I. INTRODUCTION

AUDIO-DRIVEN talking face generation has garnered significant interest in recent years due to its potential applications in multimedia applications, such as video conferences, digital humans, and virtual assistants. The system works by taking a speech sequence as input and generating a talking video synchronized with the speech content. Recently, various approaches have been proposed, leading to substantial improvements in the quality of talking face generation, including lip-sync quality [1], [2], [3], [4], per-frame visual quality [5], [6], [7], [8], and emotion controllability [9], [10]. To evaluate the quality of generated videos, existing studies typically rely on either single-frame-based metrics (e.g., PSNR, SSIM [11]) or human-centered subjective experiments.

Despite the rapid progress, one critical property of talking face videos has been largely overlooked. Before elaborating on the answer, let’s look at an example shown in Fig. 1 (top). As can be found, the two generated videos have approximately comparable image fidelity to the real in terms of per-frame quality. However, according to the user study, the baseline receives a lower score for video realness compared to the other two videos. To figure it out, on the bottom of Fig. 1, we concatenate the vertical slice of each frame for three videos to visualize the lip motion results. As can be seen, the baseline produces more jagged patterns compared to the proposed method and real videos. This leads us to pose the question: What distinguishes the talking videos between the baseline and the proposed/real versions, given that users have rated the baseline video lower in terms of video realness? We ascribe this disparity to the presence of motion jitters, which consistently result in an annoying and jittery experience in talking videos.

In this work, we delve into the motion jittering issue with systematic analyses and thereby mitigate the problem with effective solutions. We begin by selecting a basic talking face generation pipeline as the baseline model, which is widely used in previous research [2], [5], [12], [13], [14], [15], [16], [17]. In this pipeline, we first estimate the mouth-related expression parameters from the extracted audio features. Then, the expression parameters, as well as the shape and pose para-meters (from the background images) are combined as the input of a 3D face model [18] to render animated face shapes (3D face representations used in this work). We then introduce a neural renderer to synthesize realistic face images using a concatenation input of the background images and animated face shapes.
Firstly, jitters arise from 3D face representations. Although 3D face representations provide detailed mouth features and head poses for the neural renderer, the 3D representations are estimated on a frame-by-frame basis by the 3D face model, ignoring information from adjacent frames. As a result, the representations lack temporal smoothness across frames and have jitters in lip motions, which provide unstable ground truth labels for the audio-to-expression prediction network.

Second, training-inference mismatch can occur in the neural renderer. During training, the renderer is optimized to synthesize a face image from a paired background image and face shape. However, during inference, the face shape changes due to the new input audio, leading to a mismatch with the background image. When concatenating together, the model needs to generate a realistic face video with mouth shapes similar to the mouth part from the input face shape and the rest part from the background image. This introduces uncertainty for the model to handle this mismatch that has never been seen in training, thus causing motion jitters in the rendered images.

(iii) Third, the lack of dependency modeling in the neural renderer. As lip motions are naturally and inherently linked to speech, the generation of talking face videos can be represented as a sequence-to-sequence task since speech signals and videos are often viewed as sequential data. Therefore, the neural renderer must capture inherent inter-frame dependencies. However, this crucial property is frequently ignored in existing studies [5], [6], [7], [8], [12], which often learn to synthesize each image independently without regard to their dependency. As a result, they fail to model the inherent stable motions for realistic talking face videos.

To address the motion jitters in generated videos, we propose several effective solutions. (i) First, we propose to remove the jitters from the 3D face representations. A simple way for smoothing is to use moving averages or manually designed Gaussian-based smoothing kernels. However, both of them fail short in capturing lip motions with varying speeds (fast or slow), either incurring the problem of over-stable motions which eliminate the differences between similar pronunciations, or producing less motion-stable results. To combat this, we train a smoothing weight estimation network to predict adaptive weights for each frame based on the 3D face expressions. (ii) Second, to address the training-inference mismatch problem, we sought to introduce augmented erosion to background images during the training process, which simulates the potential mismatch in inference. By randomly eroding the mouth regions with different shape images, our neural renderer is more robust to the distortions in the mouth regions, thus reducing jitters. (iii) Third, we regard talking face generation as a sequence-to-sequence generation task and harness the power of the transformer [19], [20] in sequence modeling. To achieve this, we develop a transformer-based dependency modeling module, which we embed into the neural renderer. Our dependency module takes the merits of the transformer’s ability to model temporal relations, making significant contributions to improving motion stability.

Additionally, beyond addressing the issue of motion jittering itself, there is significant value in quantitatively examining the motion jitters for objective evaluations to reduce the cost of subjective experiments and assist in the design of models for talking face generation. However, evaluating motion jitters can be challenging, as they cannot be observed from a single image or two images, but rather require an image sequence, which makes it difficult to define and measure the quality of motion stability. The absence of off-the-shelf metrics that can evaluate motion stability in talking head videos hinders progress in the development of motion-stable talking face generation. To bridge the gap and facilitate future research, we develop an objective metric, namely the Motion Stability Index (MSI), which measures the motion stability in talking face videos. Specifically, we calculate the reciprocal of the variance of acceleration in sub-sequences for each facial keypoint and then take the average as an indicator of motion stability. Our experiments show that the Pearson correlation between MSI and subjective scores for motion stability is 0.490 (p < 0.001) and 0.493 (p < 0.001) for lips and jaws, respectively, indicating a statistically significant correlation between the proposed MSI and subjective scores.

Our main contributions are summarized as follows:

- We emphasize the motion jittering problem in talking face videos and provide thorough systematical analysis on motion jitters. We then explore a well-organized and effective framework to achieve motion-stable talking face generation.
- We propose a distinctive solution that encompasses several efficient designs, including an adaptive smoothing module, augmented erosion, and a transformer-based dependency modeling module.
- To facilitate future studies on motion-stable talking face generation, we introduce an effective metric, i.e., MSI, for quantifying motion stability in talking face videos.
- Extensive experiments demonstrate that our framework creates talking face videos with both high fidelity and motion stability. Furthermore, our results confirm that our metric has a statistically significant positive correlation with subjective evaluations.
II. RELATED WORK

A. Talking Face Generation

Video-Driven vs Audio-Driven: Existing works on talking face generation can be classified into three main categories: video-driven, audio-driven, and text-driven methods. Video-driven methods require an external driving source, such as another face image or talking video [21], [22], [23], [24], [25], [26], to reenact the existing facial image. Although these methods can synthesize talking face images with good visual quality, they require an additional driving video and another lip-sync audio to compose a complete talking video. On the other hand, audio-driven methods employ speech as the driving source and synthesize talking video in synchronized with the input speech. Text-driven methods [27], [28] take the text as input, but they require the synthesis of speech from text since a talking face always needs speech. Moreover, for the text-driven method, predicting the duration of text to align with the talking video is a challenge, while speech can be easily aligned with the talking video. In this paper, we focus on audio-driven talking face generation as it has more advantages over its video-driven counterparts: 1) they are more essential to talking face generation since the goal is to generate talking faces from talking audio, and 2) it does not require an additional driving video, making it generally applicable to different scenarios.

End-to-End vs Two-Stage: Various approaches have been employed to generate talking faces from audio inputs. Some works pursue an end-to-end approach, directly synthesizing images from speech features [3], [29], [30], [31], [32], [33], without requiring intermediate facial geometry to bridge the neural renderer and input speech. However, there exists an information mismatch between input audio and output video: 1) audio content has a weak correlation with the head pose changes or eyes; 2) the static features such as the shape or texture of the talking face are not in good correlation with the audio content. To overcome these limitations, some works have proposed two-stage methods that use intermediate representations to capture the target talking face, and then synthesize the talking face from the audio-predicted representations using a neural renderer. Nonetheless, since audio content has a very weak correlation with the head pose and eye gaze, generating the whole intermediate representations, i.e., lip movement, expression, head pose, and eye gaze, can lead to one-to-many mapping and ill-posed problems. Therefore, recent studies [2], [5], [9], [12], [13], [34] try to only generate the intermediate representations in the mouth area and then concatenate them with a background image to generate whole talking images. In this paper, we adopt the two-stage method as it promotes better visual quality for synthesized talking face videos.

2D Face Geometry vs 3D Face Representation: The aforementioned intermediate representations can be selected as 2D face geometry, such as facial keypoints [1], [7], [35] and parsing maps [25], [36], or 3D face representations [5], [14], [37], [38], [39]. Although using 2D face geometry as intermediate representations is straightforward, it has several drawbacks: 1) it is not fine-grained to represent a talking face with only 2D face geometry; 2) it necessitates a large number of high-quality videos for training; 3) it can cause identity leakage and face geometry deformation if 2D geometry is inaccurately predicted. Recent works generally prefer to use 3D face representations as intermediate features, since they can alleviate the above issues of 2D face geometry. In this paper, we adopt 3D face representations [3] for talking face generation but aim to address the problem of motion jitters.

B. Vision Transformer

Recently, transformer [19] has emerged as the most popular tool for modeling short- and long-term correlations in various fields [40], [41], [42]. Typically, the transformer comprises two primary components, multi-head self-attention layers, and feedforward modules. Due to its ability to model long-range relationships, transformer-based models have been applied to various vision tasks, such as visual recognition [20], 3D facial animation [42], [43], image inpainting [44], [45], object detection [46], [47], image synthesis [48], image harmonization [49], among others.

In the fields of talking face generation, previous studies have focused more on 3D meshes generation [42], [50] or facial blend-shape coefficients estimation [43], rather than photo-realistic face images. Fan et al. propose FaceFormer [42], an autoregressive transformer-based architecture that predicts moving coordinates for 3D facial mesh animation. Chen et al. [43] design a MOE-based transformer to generate facial coefficients from speech. While these methods have achieved success in animation generation, they cannot be adapted to photo-realistic talking video generation due to the high-dimensional nature of both the input and output, i.e., $\mathbb{R}^{T\times H\times W\times C}$. In this paper, we regard photo-realistic talking face generation as a frame-based sequence-to-sequence generation task and develop a transformer-based architecture to model inter-frame dependencies.

C. Motion-Stable Video Generation

To improve consistency, prior methods for talking face generation have incorporated autoregressive generation strategies [2], [13] by “looking at previously generated frames” in the neural renderer, or have included a temporal discriminator to judge the video quality on multiple consecutive frames [1], [3], [4]. Although they have made significant strides in improving frame consistency, such as consistent illumination or color [1], [2], they are not explicitly designed to improve motion stability, hence being not efficient enough in addressing the motion jittering problem (motion jitters are noticeable in their talking face videos). Yu et al. [51] utilize optical flow to warp previous images for better consistency. While achieving better temporal consistency, the optical flow only represents the motion between two adjacent frames instead of the long-range dynamics. Additionally, unstable and inaccurate flows also hinder stable motion generation.

Beyond talking face generation, several studies try to address the temporal inconsistency problems that are mainly incurred by illumination/texture/color changes between two frames [52], [53]. Unfortunately, this kind of temporal jitter of appearance
in consecutive frames is distinct from the motion jittering issue in the talking face generation, where we emphasize more about the object’s motions rather than its appearance because current talking face generation methods seldom suffer from appearance jitters. Other works aim to improve video stability for handheld captured videos due to high-frequency camera jitters during the video capturing process [54], [55], [56], [57], but not for the motion jittering problem in video synthesis.

D. Motion Stability Assessment

Unlike frame-level visual quality assessment [58], [59], [60], [61] which only evaluates the visual quality of individual frames, the evaluation of video quality encompasses not only the per-frame visual quality but also the temporal consistency of both video contents (e.g., object category and shape) and appearances (e.g., illumination, texture, and color), as well as motion stability (e.g., lip motions). To evaluate video jitters, some prior studies rely on subjective experiments to quantify the impact of frame jitters (caused by network jitters) in video transmission [62], [63], [64]. For video quality assessment, [65] proposes an NR-VQA method that enables quality assessments of the video distortions for in-the-wild videos but it is not suitable for quantifying motion jitters in talking videos. Our proposed metric aims at quantifying the motion stability in talking videos instead of arbitrary video content, bridging the gap between talking video generation and evaluation.

III. METHOD

In this section, we first introduce the overall setting of our baseline (Section III-A), followed by a systematical analysis of motion jittering problem (Section III-B) and our proposed solutions (Section III-C, Fig. 2, and Fig. 3). We then elaborate on the detailed analyses of the motion stability algorithm and our proposed metric (Section III-D). Finally, we describe our loss functions for training our networks (Section III-E). In Table I, we briefly describe the symbols defined by our method.

A. Baseline Overview

To analyze and mitigate the motion jittering problem, we investigate previous approaches that can generate personalized talking face videos and then design a baseline that is audio-driven, two-stage, and uses 3D face parameters as the intermediate representations between audio input and video output. In this section, we introduce each component of our baseline as follows.

3D Face Representations: 3D face model [18], [37] has been used to disentangle facial shape, texture, expression, and other facial properties. With a 3D face model, one can reconstruct the in-the-wild face with one-dimensional parameters vector, i.e., shape, expression, texture, and pose parameters. As a typical pipeline, taking 3D face representations to bridge the input

| Symbol | Verbal Representation |
|--------|-----------------------|
| $T$    | Frame number of an image sequence |
| $x_{1:T}$ | Sequence of eroded input face image |
| $s_{1:T}$ | Face shape sequence |
| $f_{1:T}$ | Background image sequence (ground truth) |
| $\tilde{f}_{1:T}$ | Synthesized image sequence |
| $W$ | Adaptive smoothing weight |
| $K$ | Weight size of $W$ |
| $f_{T}$ | Time-aligned audio feature sequence |
| $f_{1:T}$ | Encoder feature sequence of the neural renderer |
| $f_{2:T}$ | Input of Dependency Modeling Module |
| $f_{3:T}$ | Output of Dependency Modeling Module |
| $\tilde{s}_{1:T}$ | Smoothed face shape sequence |
| $M_{1:T}$ | Mouth mask sequence for Augmented Erosion and reconstruction loss |
| $P$ | Number of selected keypoints for 3D metric |
| $L$ | Length of sub-sequence for 3D metric |
| $\sigma$ | Constant |
| $\sigma^p(t)$ | Coordinates of the $p$-th keypoint at $t$-th frame |
| $\sigma^q(t)$ | First-order time difference of the $p$-th keypoint at $t$-th frame |
| $\sigma^q(t)$ | Second-order time difference of the $p$-th keypoint at $t$-th frame |

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
audio and output video has been validated by recent competing methods [5], [9], [12], [13], [14]. The 3D face model is used in training and inference respectively as follows. During training, we first extract 3D face parameters from the target images using DECA [18]. These parameters include facial expression, shape, and pose. These parameters serve two purposes: 1) the facial expression of the mouth area is taken as the training target of the audio2expression model in audio processing (which will be introduced later); 2) the expression, shape, and pose parameters of the whole images are used to render the face shapes, which will be combined with the background images (see Fig. 2), and then used as input for the neural renderer (introduced later).

Audio Processing: We employ a pre-trained audio features extraction model [66] to extract audio features (denoted as $f_{a_{1:T}} \in \mathbb{R}^{T \times 64}$), from which we predict the mouth-related 3D expression parameters (known as $s_{1:T} \in \mathbb{R}^{T \times 53}$, where $T$ corresponds to the number of video frames, and 53 denotes the dimension of 3D face expression parameters in our experiment) using a robust audio2expression model (Transformer-S2A [43]). By taking a sequence of audio features as input, Transformer-S2A estimates a sequence of 3D expression parameters with multiple stacked transformer layers.

Neural Renderer: The neural renderer takes the eroded background images and 3D face representations as input and generates new images frame-by-frame without modeling inter-frame dependencies. The neural renderer comprises three main components: an encoder, a bottleneck, and a decoder. The encoder consists of one convolutional layer (kernel size 7, stride 1) and 2 strided convolution layers (kernel size 4, stride 2), each of which is followed by instance normalization and leaky-relu activation. We use 8 residual blocks as [67] for the bottleneck. For the decoder, we use 2 transposed convolution layers (kernel size 4, stride 2) and 1 convolution layer (kernel size 7, stride 1) and tanh activation to produce the final images.

B. Analyzing Motion Jittering Problem

We consider the generation procedure as a sequence-to-sequence mapping problem $\hat{I}_{1:T} = G(s_{1:T}, x_{1:T}, \Theta)$, where $s_{1:T}$ and $x_{1:T}$ denote the input 3D face representations and eroded images from the time stamp of 1 to $T$, and $\Theta$ represents the network parameters. Since the neural renderer can extract visual background information (hair, backgrounds) from the eroded background images, it is easy for the neural renderer to synthesize those regions. However, the mouth and jaw parts are more challenging to deal with because 1) the original lips and jaws match the background images while the new synthesized lips and jaws should also match the background images, which introduces uncertainty to the generation process, and 2) the motions of the lip and jaw are complex and diverse due to the diverse content of input audio, which makes it more difficult to generate natural and stable motions. With systematic analyses, we identify several causes that may result in motion jitters in the synthesized videos, including $s_{1:T}$, $x_{1:T}$, and $\theta$.

First, jitters from the 3D face representations $s_{1:T}$. Despite the fact that we adopt 3D face shapes to bridge the neural renderer and audio processing module, the facial parameters are extracted frame-by-frame by the 3D face model or predicted by an audio processing module, where they can have jitters and are not stable across consecutive frames. Further, the jitters from the input parameters incur the jitters of 3D face shapes fed into the neural renderer. To this end, a straightforward solution to alleviate these jitters is to smooth the 3D face shapes in geometry space as they directly serve as the input to the neural renderer.

Second, training-inference mismatch. In training, we extract the 3D face parameters from the background images and render...
the face shapes as the input of neural renderer. In line with previous research [5], [13], [14], we remove the mouth region and concatenate the eroded image with the face shape in a channel-wise manner to create the input to the neural renderer. However, the eroded image and the face shape are from the same target image in training, which relieves the optimizing procedure. During inference, the eroded image comes from the same background video but is not complementary to the new face shape because the mouth area has been changed due to the new audio. As a result, the neural renderer must generate a realistic face with the mouth derived from the face shape while incorporating the rest from the eroded image. This challenge introduces difficulties for the neural renderer in handling the distortion, which, in turn, results in more errors around the mouth boundary.

Third, lack of inter-frame dependency modeling in the neural renderer. Each image frame in a video is not independent but correlated with its adjacent frames. However, the current neural renderer generates each frame independently, without considering information from adjacent frames. Therefore, synthesizing each frame without “looking behind and ahead” can result in inconsistent motions and motion jitters in face video. To address this issue, we propose to apply a dependency modeling module to enforce dependency learning and improve the motion stability for the current neural renderer.

C. Improving Motion Stability

1) Adaptive Smoothing: To relieve the motion jitters arising from the independent extraction of 3D face parameters, we propose to smooth the face shapes across adjacent frames. Our approach involves defining a smoothing weight as \( W \in \mathbb{R}^{T \times K} \), where \( T \) represents the total number of frames, and \( K \) is the smoothing width. For each \( t \in [1, T] \), we consider using \( W_t \in \mathbb{R}^{K} \) to smooth the face shape sequence as follows:

\[
\tilde{s}_t = \sum_{k=-(K-1)/2}^{(K-1)/2} W_{t,k} \ast s_{t+k},
\]

where \( s_{t+k} \) is the \((t+k)\)-th face shape, \( \tilde{s}_t \) is the \(t\)-th smoothed face shape.

There are various options for the smoothing weight \( W \): 1) a pre-determined, handcraft, and fixed weight; 2) a global but learnable weight (i.e., \( W_i = W_j \) for any \( i, j \in [1, T] \)), \( W \) can be optimized end-to-end; 3) an adaptive weight (i.e., \( W_i \) can be different from \( W_j \) for any \( i \neq j \)). In our experiments, we find that using handcraft and fixed weights often leads to over-smoothing that will eliminate the subtle expression variations in fast motion or less-smoothness that cannot deal with the motion jitters. Although using global and learnable weights with learned parameters that perform on the whole dataset may provide an improvement over a fixed weight, it still fails to address the issue of the fixed weight. Thus, we choose to utilize an adaptive weight approach, where each \( W_i \) is learned based on the current and adjacent frames of the input face shapes. Fig. 3 (top left) illustrates our smoothing strategy.

2) Augmented Erosion: To address the training-inference mismatch problem in the neural render and improve motion stability, we add augmented erosions on the background images during training by simulating the patterns of mismatch in inference. We employ two ways for adding augmented erosions: 1) we add random noise to the original facial expressions on the mouth area, and then create the eroded images via mouth area mask generated by [18]; 2) we randomly erode/dilate and translate/rotate the mask to simulate face deformations in inference. By incorporating this operation, the generator learns to synthesize images around the mouth region from different image-shape patterns, thus becoming more robust in the inference stage. The results of augmented erosion can be seen in Fig. 3 (top right).

3) Transformer-Based Spatial-Temporal Dependency Modeling: Like natural language or speech, videos constitute sequential data comprised of multiple consecutive frames. To mitigate the motion jitters in the synthesized video, dependency modeling must be taken into consideration. Recently, transformer-based methods have outperformed RNN-based methods in handling sequential data by eschewing recurrent modeling schemes but relying on a self-attention mechanism. To explicitly model inter-frame dependencies and generate talking face videos in a sequence-to-sequence generation manner, in this work, we design a transformer-based dependency modeling module that employs several stacked Transformer blocks in the convolutional encoder-decoder architecture.

As depicted in Fig. 3 (bottom left), the Dependency Modeling Module takes the fused features \( f_{1:T} \) as input, where \( f_i \in \mathbb{R}^{h \times w \times c}, h = w = 64, c = 256 \). To formulate the sequence embedding for the transformer, we first split each fused feature into \( p \times p \) smaller patches for each sample \( f_i \). Then all patches (each patch has shape \( 64/p \times 64/p \times c \)) will be flattened and linearly projected into tokens \( z \in \mathbb{R}^{(T \cdot p \cdot p) \times d} \), where \( d = 512 \). Similar to ViT [20], our transformer encoder is stacked by \( N \) transformer blocks, each consisting of two main components: the spatial-temporal self-attention mechanism and the feedforward layers. By taking the sequential data of projected patches as input, the dependency modeling module performs self-attention and feature learning to model temporal and spatial dependencies. The output of our transformer block will be linearly projected and reshaped to \( p \times p \) patches, which are then composed and fed into a convolutional decoder to recover the facial details from the output features \( f'_{1:T} \). In our experiments, we set \( p = 4 \), and \( N = 4 \).

Compared to the baseline that independently synthesizes each face image, our transformer-based neural renderer is capable of learning inter-frame dependency and synthesizing talking face videos with better inherent face dynamics. Compared to RNN-based models, our transformer-based neural render can not only capture both the spatial and temporal dependencies conditioning on the whole sequence via self-attention mechanism but also enable the training and inference in parallel, which introduces less latency than RNN-based methods. Benefiting from the transformer-based architecture, the neural renderer is capable of creating talking face videos with more consistent and stable motion.

In addition, considering the issue of information lost about the tongue and teeth in 3D face representations, we design an Audio Fusion Module to fuse the audio and visual features before
dependency modeling. We employ 4 convolution-based residual blocks for audio fusion, and the detailed structure is shown in Fig. 3 (bottom right).

D. Motion Stability Index (MSI)

As discussed in Section I, quantitatively evaluating the motion stability of talking face videos is challenging and valuable. Yet, for previous works such as [2], [5], [6], [7], no off-the-shelf objective metric is available for them to measure the motion stability, thus necessitating reliance on subjective user studies. To bridge this gap, we endeavor to develop an effective metric that can objectively quantify motion stability in talking face videos.

1) Analysis: To achieve this, two fundamental questions need to be answered: 1) Which specific aspect of motion should be measured; and 2) Which method is effective to measure motion stability in these videos? We attempt to answer these questions from two perspectives:

(i) To represent motions in talking videos, an intuitive choice is optical flow [68] which represents the motion of each pixel between two frames. However, it has some limitations that make it unsuitable for motion descriptions in talking videos. First, many pixels in the first frame will disappear in the subsequent frames due to occlusions and head movements, making it difficult to formulate complete motion trajectories in a video. Second, the lack of semantic labels in optical flow makes it impossible to locate corresponding pixels in different frames. Additionally, optical flow is sensitive to illumination changes [69]. These limitations make optical flow not suitable for representing motion sequences in talking videos. In contrast, we propose utilizing facial keypoints to represent the essential motions of talking faces. This approach has two notable advantages: Firstly, facial keypoints are semantically meaningful, facilitating the description of motion. Secondly, it is robust to occlusions and lip motions. By utilizing the coordinates of each keypoint, we formulate a complete motion sequence and obtain the trajectories of all selected keypoints.

(ii) Motion stability in talking face videos is an abstract concept that is difficult to decompose and quantify. Therefore, we sought to consider motion jitters as an alternative as they are inversely related to motion stability. Meanwhile, motion jitters can reveal high-frequency and irregular changes in motion throughout a video, and a higher level of motion jitters implies a lower level of motion stability. To quantify the jitter level of a keypoint trajectory, one can calculate the variance of its motion or accelerations. Similarly, a smaller variance indicates more stable motion, while a larger variance reflects more severe motion jitters.

2) Our Solution: Driven by these analyses, we propose an intuitive yet effective metric, i.e., motion stability index (MSI), to assess motion stability for talking face videos. Our algorithm (shown in Algorithm 1), comprises two key steps: (i) We use facial keypoints (selected from the lip and jaw regions) to accurately compute facial motions in each video, and these keypoints can be accurately regressed through a deep-learned face alignment tool [70]. With the regressed facial keypoints, we can formulate a coordinate sequence to compute the motion at each frame using first-order difference. (ii) We sought to measure the motion jitters by computing the variance of motion accelerations. However, we recognize that the variance may be influenced by the length of the target sequence. To address this, we partition the target sequence into sub-sequences, each with the same length, and then average the scores of all sub-sequences as the motion stability of the target sequence. Our solution is grounded on the assumption that the number of frames varies from video to video. Hence, using different sequence lengths for both long videos and short videos is neither fair nor wise. Instead, we equally regard each sub-sequence as an independent sequence, regardless of whether it is derived from a long or short video.

3) Algorithm: In the first step, we extract all image frames from the talking face video and obtain facial keypoints as a sequence. We focus on the keypoints on the lip and jaw since the mouth and jaws are the main components driven by audio. It is worth mentioning that the MSI scores in the lip region differ from those in the jaw region due to their different motion scales. Therefore, it is preferable to measure MSI separately for the lip and jaw. As shown in Fig. 4(a), we sample 20 and 11 keypoints for the lip and jaw, respectively. For the i-th point in the t-th frame, we extract all image frames and obtain facial keypoints.

Algorithm 1: Keypoint-Based Motion Stability Index.

Input: An image sequence \( \{I_1, I_2, \ldots, I_T\} \), the sub-sequence length \( L \), constant \( C \), \( P \) selected keypoints.

1: Regress facial keypoints’ coordinates \( Z_{1:T}^{1:P} \) for \( \{I_1, I_2, \ldots, I_T\} \).

2: for \( i = 1, \ldots, P \) do

3: Query the \( i \)-th keypoint’s coordinates \( \{Z_{1:T}^1, Z_{1:T}^2, \ldots, Z_{1:T}^P\} \) from \( Z_{1:T}^{1:P} \).

4: Generate sequence \( \{a_1^i, a_2^i, \ldots, a_T^i\} \) according to (2).

5: for \( t = \frac{1}{2}, \frac{3}{2}, \ldots, T - \frac{1}{2} \) do

6: Generate a sub-sequence \( \{a_{t-\frac{1}{2}}, a_{t-\frac{1}{2}+1}, \ldots, a_{t+\frac{1}{2}}\} \).

7: Update \( a_t^i \) according to (3).

8: end for

9: Update \( MSI^C(I_{1:T}) \) according to (4).

10: end for

Output: \( MSI(I_{1:T}) \)
frame, we denote its location as $Z_t^i \in \mathbb{R}^2$. Further, we compute the first-order time difference of keypoint coordinates to obtain the motions $v_t^i$. Similarly, we compute the second-order time difference as acceleration $a_t^i$:

$$v_t^i = Z_t^i - Z_{t-1}^i, \quad a_t^i = v_t^i - v_{t-1}^i, \quad t = 1, 2, \ldots, T,$$

where $v_0^i = 0$.

In the second step, the acceleration sequence is partitioned into a series of contiguous sub-sequences using a sliding window of length $L$. As demonstrated in Fig. 4(b), the window is moved forward in increments of $L/2$ at each step, generating $2T/L$ sub-sequences. Therefore, we assess the motion jitter within a short sub-sequence through statistical variance:

$$\sigma_t^i(a) = \text{Var}(a_t^i - \frac{1}{L} \sum_{1 \leq t \leq L} a_t^i),$$

where $t \in \mathcal{T} = \{ \frac{L}{2}, \ldots, T - \frac{L}{2} \}$. However, the variance calculated from the acceleration sequence of a talking video might span a wide range, e.g., $\sigma_t^i \in [0, 20]$. Moreover, the motion jitter is inversely correlated with motion stability, which is not intuitive for human understanding. Therefore, we use the modified reciprocal of $\sigma_t^i$ to represent the stability score of motions associated with the $i$-th keypoint:

$$\text{MSI}^i(I_{1:T}) = \frac{L}{2T} \sum_{t \in \mathcal{T}} \frac{1}{\sigma_t^i + C},$$

where the constant $C$ is included to avoid instability when $\sigma_t^i$ is close to zero. For example, $\sigma_t^i$ might be very small when the mouth is closed during pauses.

By measuring the motion stability index $\text{MSI}^i(I_{1:T})$ for the selected keypoints, we formulate the average MSI score:

$$\text{MSI}(I_{1:T}) = \frac{1}{P} \sum_{i=1}^{P} \text{MSI}^i(I_{1:T}).$$

In practice, we observe that the size of the face can scale keypoints’ coordinates and introduce unwanted perturbations to the final results. To eliminate these impacts, we adopt a unified strategy to crop the video with a fixed bounding box for all test videos. Specifically, we use the first 10 frames to regress a mean bounding box, and then rescale the box to ensure that the ratio of the mouth width to the box width is 0.25. Furthermore, for a fair comparison, we rescale the cropped face videos into resolution $256 \times 256$ for all methods.

E. Loss Functions

Given paired audio features, our neural renderer is trained in a supervised manner to generate a sequence of realistic images that closely resemble the ground truth. We used two kinds of losses, namely reconstruction loss (including unbalanced pixel-wise re-

$$\mathcal{L}_{\text{adv}} = -\mathbb{E}_{t \in [1, T]}(\log D(I_t)) - \mathbb{E}_{t \in [1, T]}(\log (1 - D(\hat{I}_t))),$$

where $D$ shares the same structure as [5]. The full objective of the neural renderer is:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{rec}} + \lambda_2 \mathcal{L}_{\text{adv}},$$

where $\lambda_1$ and $\lambda_2$ are set to 20 and 1, respectively.

IV. EXPERIMENTS

A. Experimental Setup

1) Dataset and Preprocessing: Our method aims at synthesizing personalized talking face videos driven by audio, achieving high-fidelity generation results with only a short video of a few minutes. Therefore, we prepare several videos of different individuals for our experimentation purposes. Specifically, we collect two weekly speech videos of Obama from [6], [7] (one approximately 3 minutes long and another around 4 minutes long), with 85%/5%/10% used for training/validation/testing, respectively. To test our method on standard lip reading datasets, we conduct experiments on the GRID [72] dataset, which contains 50 minutes of videos for each person. Following LipSync3D [2], we use 50%/20%/30% of the corpus for training/validation/testing, respectively.

2) Implementation Details: We implement our system with two detailed pipelines: audio2expression and neural rendering. For the audio2expression predicting pipeline, we extract audio features from the audio track and 3D facial expression parameters from the corresponding video frames. The audio features and 3D facial expressions from the same timestamp are viewed as paired training samples. Since we focus on improving the motion stability of talking face videos, we directly use the existing Transformer-S2A [43] and follow their settings to train the audio2expression model. In the second stage, we treat consecutive frames as an input sequence and train the neural renderer and adaptive smoothing module. Initially, we optimize the neural renderer without the adaptive smoothing module for 100 epochs and then jointly train them together for another 20 epochs. Since the adaptive smoothing module is lightweight, we optimize it using a smaller learning rate (1e-6) than the neural renderer (1e-4). During training, we apply augmented erosions to the input images and obtain the eroded images, which we concatenate with the smoothed face shapes as the input of the neural renderer.
We compare our approach with recent representative works: NVP [5], LipSync3D [2], LiveSP [7], and AD-NeRF [6]. In particular, we retrain the AD-NeRF [6] model on our dataset and conduct testing. We use the released code and pre-trained checkpoint of LiveSP without modification. As NVP is not open-sourced, we implemented the neural rendering network for NVP and trained the rendering model on our dataset, with the audio2expression network being the same as ours. LipSync3D neither makes its code available nor elaborates on detailed network structures. Therefore, we present the comparison results on the GRID [72] dataset and adopt the quantitative results that are reported in their original paper.

2) First-Order Vs. Second-Order: We also compare the correlations between subjective scores and different objectives for variance calculation, i.e., using either first-order time difference \( \{v_i\} \) or the second-order \( \{v_i^2\} \) for (3). We present the correlations in Table IV (the 2nd column and 3rd column). It can be found that the correlations between MSI metric and motion stability in videos reach 0.490 for the lip and 0.493 for the jaw, surpassing the effectiveness of the first-order differential measure.

3) Variance Vs. Modified Reciprocal of Variance: Instead of measuring the variance of second-order time difference across frames, we calculate the average of their modified reciprocal values to achieve higher correlations. As shown in Table IV (the 3rd column and 4th column), our modified reciprocal of variance is better than the original variance measurement in terms of correlation.

4) Discussion: Our MSI has several advantages. First, with accurate keypoint regression for all frames, we can formulate robust motion trajectories for lips and jaws, making it effective in assessing the motion stability in talking face videos. Moreover, it is robust to the video length and capable of reporting an average score of motion stability across all video segments. Although it cannot be used as a versatile metric of the overall generation quality of talking videos, our metric can serve as an alternative objective metric to subjective experiments in assessing motion stability, thus benefiting future studies on talking head generation.

### B. Efficacy of MSI

Previous talking face generation methods have to rely on subjective user studies to evaluate the quality of the generated videos. In this work, we propose an effective new metric called MSI, to fill the blank in the talking face generation field. To validate the efficacy of MSI, we conduct subjective experiments to score motion stability on a series of talking head videos generated by different approaches, and then calculate the Pearson correlation coefficients between the subjective scores and MSI. As described in Section III-D, we split the keypoints into two groups and discuss the results for lip and jaw regions separately.

1) Impact of \( L \) and \( C \): We first investigate the influence of sub-sequence length \( L \) (discrete number) and constant \( C \) (same) on the correlation coefficient. On the one hand, since there is no motion jitter between two images, we explore the data range of \( L \) from 3 to the whole sequence length \( T \) by enumerating some typical discrete numbers. On the other hand, \( C \) is used to avoid instability when \( \sigma_i \) is close to zero but larger \( C \) narrows the difference of different variances while smaller values of \( C \) fail to avoid the negative impacts of variances that are close to zero. We provide some representative results based on different \((L, C)\) pairs and show them in Table II (MSI-Lip) and Table III (MSI-Jaw). As can be inferred, using sequences that are too short or too long leads to unsatisfactory results. For example, the correlations are almost 0.42 (0.44) for MSI-Lip (MSI-Jaw) if we focus on the whole sequence. We highlight the relatively high correlation results with a gray background. As can be found, the correlation for \( L \) is in the range of \((5, 12)\). The results of \( L = 12 \) and \( C = 0.5 \) achieve high correlations for both MSI-Lip and MSI-Jaw, i.e., 0.490 and 0.493, which are selected as our default settings. Our experimental results on this task suggest that evaluating the motion jitters in different talking-face videos is feasible through the same metric, albeit with room for improvement.

### C. Comparisons With State-of-The-Arts

1) Comparison Baselines: We compare our approach with recent representative works: NVP [5], LipSync3D [2], LiveSP [7], and AD-NeRF [6]. In particular, we retrain the AD-NeRF [6] model on our dataset and conduct testing. We use the released code and pre-trained checkpoint of LiveSP without modification. As NVP is not open-sourced, we implemented the neural rendering network for NVP and trained the rendering model on our dataset, with the audio2expression network being the same as ours. LipSync3D neither makes its code available nor elaborates on detailed network structures. Therefore, we present the comparison results on the GRID [72] dataset and adopt the quantitative results that are reported in their original paper.

2) Evaluation Metrics: It is worth noting that no single metric can provide a comprehensive evaluation of overall video quality because video quality encompasses various facets, including but not limited to motion stability/jitters, aesthetics quality of each frame, pixel-wise errors, shape similarity, and image fidelity. Therefore, we rely on several metrics together as well as subjective user studies to assess the generation quality.
from various perspectives. In particular, we adopt the following metrics:

- **SSIM** (Structural SIMilarity) [11] a full-reference metric that is capable of measuring the structural similarity between the generated image and the ground truth.
- **CPBD** (Cumulative Probability Blur Detection) [73], a no-reference metric that is employed to evaluate the sharpness of given images.
- **NLMD** (Normalized LandMark Distance), a reduced-reference metric that assesses the shape similarity of lips between the generated and ground truth. It removes the influence of different input image resolutions by normalizing the landmark distance [1] with image resolution.
- **Sync-C** (Syncnet Confidence) is a no-reference metric that measures the lip-sync score via SyncNet [74]. A higher score means better audio-visual synchronization accuracy.
- **MSI** (Motion Stability Index) is a no-reference metric that evaluates the motion stability of the lips and jaws in talking head videos.

We compute the results of SSIM, CPBD, and NLMD on generated videos frame-by-frame. As for Sync-C and MSI, we measure the metric results on the whole video.

3) **Comparison Settings:** We compare our method with state-of-the-art methods under two different settings: 1) **Paired audio-driven setting** where the audios are extracted from ground truth videos. We employ their released pre-trained models to test or retrain their models on our dataset. For quantitative evaluation, we calculate the results of all metrics on generated videos. 2) **Unpaired audio-driven setting** where the driving audios are randomly extracted from different videos. Note that the audio and original videos are not matched (without ground truth videos), we can only perform comparisons using no-reference metrics, i.e., CPBD, Sync-C, and MSI. Besides, we conduct subjective experiments since subjective evaluation can reflect the video quality from human perspectives. The synthesized videos of NVP, LipSync3D, LiveSP, and AD-NeRF are all taken from their web demos, which represents the best performance of these methods.

4) **Quantitative Evaluation Results:** The quantitative results obtained under the paired audio-driven setting are shown in Table V, from which it can be concluded that our method outperforms state-of-the-art methods in most metrics. As shown by the results of the Obama test videos, our approach successfully generates fine-grained talking face videos that exhibit good structure similarity and accurate lip-sync quality. NVP and AD-NeRF yield lower scores concerning the lip-sync quality and motion stability of lips than other methods. This is perhaps because the renderers in NVP [5] and AD-NeRF [6] synthesize image frame-by-frame, without taking into account the benefits of audio features input and dependency modeling (compared to our method). Note that the pre-trained LiveSP model is trained on videos with $512 \times 512$ resolution, which makes their result on CPBD better than ours. Overall, our approach shows the best motion stability and lip-sync quality while exhibiting fewer pixel errors and landmark distance. In comparison with LipSync3D [2], we report quantitative results on the GRID [72]. As shown in Table V (bottom), our method consistently outperforms LipSync3D in terms of SSIM, CPBD, Sync-C, and MSI-Lip, demonstrating better performance in terms of sharpness, lip-sync quality, and motion stability.

We proceed to take a closer look at model performance in an unpaired audio-driven setting. By explicitly learning adaptive smoothing weight and dependency modeling, our model effectively generates talking videos with both stable and accurate lip motions. The quantitative results are shown in Table VI. For CPBD, we observe that our method is slightly behind LiveSP, which should be ascribed to the advantages of high-resolution training data that [7] used. However, it is important to highlight that our method achieves notably higher scores in MSI than LiveSP while exhibiting significantly superior motion stability when compared to other methods. This demonstrates the effectiveness of our solutions in addressing the challenge of motion jittering. Furthermore, our performance on Sync-C is the best among all methods, showing the superior capability of our talking head generation system.

5) **User Study:** Different from previous works that only evaluated visual frame quality and lip-sync quality, we introduce a new evaluating criterion, namely motion stability, to our study. Our primary evaluation rule is the mean opinion score (MOS) given by 18 experienced users who are invited to grade the videos based on the perceived quality of motion stability, lip-sync quality, and video realness. In the user study, we ask three questions pertaining to motion stability, lip-sync accuracy, and video realism:

a) **Motion Stability:** Are the motions around the lips and jaw stable across frames?

b) **Lip-sync Quality:** How does the motion of the lips match the speech in lip-audio accuracy?

c) **Video Realness:** How would you rate the video’s realness?
Fig. 5. Visual comparison of AD-NeRF [6], LiveSP [7], NVP [5], and our method. Results of LiveSP and NVP are cropped from their web demos. We highlight some representative visual artifacts with colored arrows. For example, Red arrows point out the blurry artifacts, in consistent jaw boundaries, or eyebrows. Blue arrows reveal the inaccurate lip motions in LiveSP. Green arrows highlight the nearly muted lips generated by NVP.

### TABLE VII

**User Study Results**

| Method    | Motion Stability | Lip-Sync. Qual ↑ | Video Real ↑ |
|-----------|-----------------|-----------------|--------------|
| NVP [5]   | 3.14±0.26       | 2.63±0.22       | 2.95±0.23    |
| LipSync3D [2] | 3.05±0.27     | 3.20±0.28       | 3.01±0.22    |
| LiveSP [7] | 2.47±0.26       | 2.56±0.27       | 2.61±0.23    |
| AD-NeRF [6] | 2.54±0.30     | 3.15±0.27       | 2.65±0.27    |
| Ours      | 3.72±0.22       | 3.60±0.21       | 3.68±0.20    |

We present the average scores of MOS on motion stability, lip-sync quality, and video realness. The best/second-best results are marked in bold/underline.

The volunteers are required to rate the videos on a five-point scale ranging from Very Poor (1) to Very Good (5). A detailed annotation guideline and several real/fake videos are provided to ensure users can fully understand different grades.

Our unique framework and effective modules enable us to achieve remarkable results and outperform prior arts by a significant margin. Table VII presents the subjective MOS scores. It can be found from the table that NVP achieves the second-best in motion stability but it is not competitive in lip-sync quality compared to other approaches. This is because the amplitudes of lip motions in videos of NVP are small, which leads to a lower score in lip-sync quality. LiveSP [7] is rated with a mean score of 2.47 in motion stability and the lowest score in lip-sync quality since their generated lip motions are too rigid and have severe motion jitters. Although AD-NeRF [6] claims its capability to generate talking videos with different poses, it produces artifacts of cracks between the torso and head, which compromises its overall quality. While LipSync3D [2] attains the second-best performance in lip-sync quality and video realness, it falls far behind our approach in terms of overall performance.

6) **Qualitative Evaluation Results:** To compare the generated videos of each method, we extract key frames from their respective videos and present the visual comparison results in Fig. 5. For the images synthesized by AD-NeRF [6], some of them suffer from blurry artifacts (first column and third column), or clear cracks between the head and torso (third column and sixth column), or severe motion jitters for the torso, which also degrades their perceived realness. NVP [5] generates videos with nearly muted lip movements. Such visual results also conform to the low Sync-C and Motion Stability score in Table VI and Table VII, respectively. Despite LiveSP’s [7] advantage in producing sharp details, it still falls short in generating accurate lip motions that are synchronized with the audio. Thanks to our effective framework designs, our approach gets rid of the drawbacks of these methods. It is capable of generating plausible results, performing competitively in terms of accurate yet stable motions, satisfactory per-frame image fidelity, and fewer artifacts. We highly recommend readers to watch the videos for a better viewing experience.

D. **Ablation Study**

To fully demonstrate the efficacy of our proposed designs, we conduct extensive ablation studies on four distinct components, namely adaptive smoothing, augmented erosion, dependency modeling, and audio fusion module.

1) **Quantitative Ablation Results:** The quantitative results are presented in Table VIII, where we report the ablation results on all objective metrics. For each experiment, we remove one
TABLE VIII
ABLACTION STUDIES ON EACH COMPONENT IN OUR PROPOSED FRAMEWORK

| Method                        | SSIM↑ | CPBD↑ | Sync-C↑ | NLMD↓ | MSI-Lip↑ | MSI-Jaw↑ |
|-------------------------------|-------|-------|---------|-------|----------|----------|
| Full Model                    | 0.952 | 0.282 | 5.54    | 0.0103| 0.563    | 0.648    |
| w/o Adaptive Smoothing        | 0.948 | 0.278 | 5.54    | 0.0111| 0.491    | 0.587    |
| w/o Augmented Erosion         | 0.943 | 0.285 | 5.50    | 0.0111| 0.545    | 0.639    |
| w/o Dependency Modeling       | 0.938 | 0.279 | 5.04    | 0.0125| 0.506    | 0.619    |
| w/o Audio Fusion Module       | 0.943 | 0.281 | 5.32    | 0.0118| 0.550    | 0.636    |
| Baseline Model                | 0.939 | 0.275 | 5.03    | 0.0126| 0.439    | 0.547    |

Fig. 6. Ablation study on audio fusion module.

key module from the full model to investigate its contribution. From the results, we conclude that each module contributes to the overall generation quality of our method. Our dependency modeling module particularly helps to enhance lip-sync quality (Sync-C), lip shape similarity (NLMD), and motion stability, with higher scores of MSI-Lip and MSI-Jaw in comparison to the model w/o Dependency Modeling. The full model w/o augmented erosion (we keep the position of the mask consistent rather than changing it randomly) has higher CPBD results but is less competitive in reconstruction quality and motion stability. Meanwhile, for image similarity and lip shape similarity, our full model exhibits obvious improvements (higher SSIM and lower NLMD) over the method without augmented erosion or adaptive smoothing module. It can also be observed that SSIM results deteriorate after removing each of our proposed components or all of them. All of these observations validate the effectiveness of each design in our framework.

2) **Qualitative Ablation Results:** In Fig. 6, we showcase some visual results of the model with and without the audio fusion module. As can be found, when the audio fusion module is removed from the neural renderer, it may produce inaccurate lips which can be different from the full model and ground truth. The reason is that part of the information has been lost during the mapping of audio to facial expressions. For example, the 3D face meshes cannot preserve the details of the tongue and teeth. When fed with audio features, our neural renderer can better capture details of talking faces and generate more accurate lips.

In Fig. 7, we concatenate the vertical slice in each frame along the time for each video to show the motion jittering patterns. The position and size of the vertical slice are shown in the first row and kept the same for all videos. We have two observations. First, without each component, one can observe some jittering patterns from the vertical slice figure. In contrast, the vertical slice figure of the full model is much smoother and has significant improvements over the baseline. Meanwhile, the jagged patterns can also be observed from other videos without each component, although it may not be very clear. Second, it may be concluded from the figure that the jitters are more likely to occur in the lip regions (green box) than the nose region (red box) as the lip moves more frequently than the nose.

3) **User Preference:** In Fig. 8, we conduct user preference experiments results in terms of motion stability, lip-sync quality, and realness. From each sub-figure, it can be observed that the user preference for motion stability decreases when we remove the adaptive smoothing module, dependency modeling, or augmented erosion. Fig. 9 shows user preference results on different smoothing strategies, e.g., fixed smoothing weight, global/learnable weight, or our adaptive smoothing weight. We can infer from the results that our adaptive smoothing choice outperforms the fixed weight or global/learnable weight in lip-sync quality and video realness since the adaptive smoothing
strategy attains more votes than the rest. On the contrary, it is hard for both fixed weight and global/learnable weight to find a trade-off between both smooth quality and lip-sync quality.

4) Impact of Keypoint Accuracy on MSI Correlations: We investigate the impact of keypoint accuracy using the 2D-FAN tool [75] (which performs worse than [70]) and subsequently compute the correlation between MSI and subjective scores. The results indicate that employing a more precise tool presents a stronger correlation. Specifically, when utilizing the 2D-FAN to estimate facial keypoints, the computed MSI metrics exhibited lower correlations of only 0.283 (MSI-Lip) and 0.306 (MSI-Jaw). This finding aligns with the observation that the accuracy of keypoint detection significantly affects the reliability of the MSI metric.

5) Incorporating MSI Into Audio2Expression Training: Due to the non-differentiable nature of the keypoint detection tool [70], we cannot measure differentiable MSI loss for the training of the neural rendering model. Instead, we investigate the impact of keypoint accuracy using the 2D-FAN tool that performs worse than [70]. We subsequently compute the correlation between MSI and subjective scores. The results indicate that employing a more precise tool presents a stronger correlation. Specifically, when utilizing the 2D-FAN to estimate facial keypoints, the computed MSI metrics exhibited lower correlations of only 0.283 (MSI-Lip) and 0.306 (MSI-Jaw). This finding aligns with the observation that the accuracy of keypoint detection significantly affects the reliability of the MSI metric.

In addition to the aforementioned results in Section IV-C, we also compare our approach to Vougioukas et al. [4] and Wav2Lip [3]. Wav2Lip uses SyncNet [74] for pretraining, which implicitly incorporates temporal information by introducing a sequence discriminator or predicting the lip-sync score of the generated frames. Vougioukas et al. introduces a sequence discriminator to evaluate the realism of the generated images. For both methods, we adopt their pre-trained model1 or their project website2 to perform the test. In Fig. 10, we present qualitative examples of three methods, i.e., Vougioukas et al., Wav2Lip, and ours. As can be found, Vougioukas et al. generates talking faces with severe image distortion and mosaic effects, which significantly compromise the video quality and user experience. Compared to Vougioukas et al., Wav2Lip produces lip motions that better match the audio content but it suffers from blurry mouth artifacts (visible in almost all frames) and double-chin problems (indicated by the green arrow). On the other hand, our approach can synthesize images with both superior visual quality and plausible teeth details, demonstrating the remarkable improvement of our proposed framework on competing methods.

Table IX, as a supplement to qualitative comparison, offers quantitative comparison results with the aforementioned two methods. Notably, the distorted faces produced by Vougioukas et al. make it impossible to detect complete facial keypoints with face detection tools, thus we cannot measure the results of NLMd, MSI-Lip, and MSI-Jaw for this method. As shown in Table IX, our approach achieves higher SSIM, CPBD, and NLMd compared to the other two methods. Compared to Wav2Lip, our approach claims strengths in terms of MSI-Lip and MSI-Jaw, as it not only explicitly learns dependencies in the neural renderer but also benefits from our adaptive smoothing strategy and augmented erosion. On the other hand, although Wav2Lip has a higher Sync-C score, it requires long-time training on a larger training corpus than our approach. In contrast, our rendering module requires only three minutes of personal talking videos to achieve comparable results. Additionally, both Wav2Lip and Vougioukas et al. utilize sequence discriminators

---

1[Online]. Available: https://github.com/DinoMan/speech-driven-animation
2[Online]. Available: https://bhaasha.iiit.ac.in/lipsync

### Table IX

| Method          | SSIM† | CPBD† | Sync-C† | NLMd△ | MSI-Lip/Jaw△ |
|-----------------|-------|-------|---------|-------|--------------|
| Vougioukas et al. | 0.503 | 0.057 | 5.55    | -/2   | -/2          |
| Wav2Lip         | 0.992 | 0.211 | 6.98    | 0.912 | 0.447/0.453  |
| Ours            | 0.921 | 0.278 | 6.24    | 0.909 | 0.735/0.646  |

(MSI-Lip/Jaw: 0.585/0.677) when compared to the full model (Lip/Jaw: 0.563/0.648). On the other hand, since the MSI loss encourages the model to reduce motion variation, it simultaneously diminishes the range and diversity of mouth movements, resulting in a reduction in lip-sync quality.

### E. Additional Comparisons to Other Methods

In addition to the aforementioned results in Section IV-C, we also compare our approach to Vougioukas et al. [4] and Wav2Lip [3]. Wav2Lip uses SyncNet [74] for pretraining, which implicitly incorporates temporal information by introducing a sequence discriminator or predicting the lip-sync score of the generated frames. Vougioukas et al. introduces a sequence discriminator to evaluate the realism of the generated images. For both methods, we adopt their pre-trained model1 or their project website2 to perform the test. In Fig. 10, we present qualitative examples of three methods, i.e., Vougioukas et al., Wav2Lip, and ours. As can be found, Vougioukas et al. generates talking faces with severe image distortion and mosaic effects, which significantly compromise the video quality and user experience. Compared to Vougioukas et al., Wav2Lip produces lip motions that better match the audio content but it suffers from blurry mouth artifacts (visible in almost all frames) and double-chin problems (indicated by the green arrow). On the other hand, our approach can synthesize images with both superior visual quality and plausible teeth details, demonstrating the remarkable improvement of our proposed framework on competing methods.

Table IX, as a supplement to qualitative comparison, offers quantitative comparison results with the aforementioned two methods. Notably, the distorted faces produced by Vougioukas et al. make it impossible to detect complete facial keypoints with face detection tools, thus we cannot measure the results of NLMd, MSI-Lip, and MSI-Jaw for this method. As shown in Table IX, our approach achieves higher SSIM, CPBD, and NLMd compared to the other two methods. Compared to Wav2Lip, our approach claims strengths in terms of MSI-Lip and MSI-Jaw, as it not only explicitly learns dependencies in the neural renderer but also benefits from our adaptive smoothing strategy and augmented erosion. On the other hand, although Wav2Lip has a higher Sync-C score, it requires long-time training on a larger training corpus than our approach. In contrast, our rendering module requires only three minutes of personal talking videos to achieve comparable results. Additionally, both Wav2Lip and Vougioukas et al. utilize sequence discriminators

---

1[Online]. Available: https://github.com/DinoMan/speech-driven-animation
2[Online]. Available: https://bhaasha.iiit.ac.in/lipsync

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
and add adversarial losses on the whole sequence, but their generators synthesize face images frame-by-frame, leading to more motion jitters. In our experiment, we observe no improvements when incorporating a sequence discriminator for training the neural rendering model.

V. DISCUSSION

A. Potential Benefits

Our technology of photo-realistic talking face generation aims for goodness, value, and productivity in multimedia applications. For instance, the application of video re-dubbing has a potentially beneficial promotion in online education in different countries and regions, no matter what the original spoken language is. Our technology is also capable of synthesizing news broadcasting videos under supervision, which can greatly reduce human efforts in newscasting. In addition, the technique could also be used to synthesize motion-stable talking videos as negative samples of deepfake detection [77].

We also introduce a novel view that considers the video quality from the perspective of motion stability and motion trajectories. Our proposed objective metric suggests a close correlation between MSI and subjective human evaluation, which indicates that objective metric MSI and subjective human evaluation can complement each other in predicting or explaining one another. This discovery means our work has the potential to promote future research in this area.

B. Ethics Consideration

Although these techniques offer potential benefits, there are also ethical risks associated with their use, as they could be misused and cause harm. Therefore, we strongly recommend that the usage of our technology be closely monitored and any videos generated by our method should be clearly labeled with visible watermarks. Additionally, requests for code would be carefully assessed to ensure responsible use. We believe that with the joint efforts of the research community and the industry, we could find a promising way to foster the technology while keeping Pandora’s box close.

VI. CONCLUSION

Research in talking face generation has devoted much less attention to motion stability than aspects of per-frame visual quality or lip-sync quality, while we addressed this issue and developed a motion-stable talking face generation system. In this work, we systematically analyze the problem of motion jitters in talking face videos and propose three key solutions to address this issue. Moreover, considering no off-the-shelf objective metric is available to assess motion stability, we propose an effective metric (MSI) to bridge that gap. Our metric has close interactions with subjective evaluations, showing a promising positive correlation with subjective human evaluation. To our knowledge, MSI is the first objective metric that is capable of effectively evaluating the motion stability in talking face videos. Extensive experiments, including both objective and subjective evaluations, are conducted to demonstrate the effectiveness of our work in improving and assessing the motion stability for talking face videos. For future work, we will extend to more application scenarios, such as text-prompted emotional talking face generation, which is more challenging for motion stability but also makes talking face video more expressive and controllable.

REFERENCES

[1] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, “Hierarchical cross-modal talking face generation with dynamic pixel-wise loss,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 7832–7841.
[2] A. Lahiri, V. Kwatra, C. Frueh, J. Lewis, and C. Bregler, “LipSync3D: Data-efficient learning of personalized 3D talking faces from video using pose and lighting normalization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 2755–2764.
[3] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, “A lip sync expert is all you need for speech to lip generation in the wild;” in Proc. ACM Int. Conf. Multimedia, 2020, pp. 484–492.
[4] K. Vougioukas, S. Petridis, and M. Pantic, “Realistic speech-driven facial animation with gans,” Int. J. Comput. Vis., vol. 128, no. 5, pp. 1398–1413, 2020.
[5] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Niebner, “Neural voice puppetry: Audio-driven facial reenactment,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 716–731.
[6] 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. Int. Conf. Comput. Vis., 2021, pp. 5784–5794.
G. Zhai and X. Min, “Perceptual image quality assessment: A survey,” Sci. China Inf. Sci., vol. 63, pp. 1–52, 2020.

A. Chetouani and M. Pedersen, “Image quality assessment without reference by combining deep learning-based features and viewing distance,” Appl. Sci., vol. 11, no. 10, 2021, Art. no. 4661.

Q. Huynh-Thu and M. Ghanbari, “Impact of jitter and jerkiness on perceived video quality,” in Proc. Worksh. Video Process. Qual. Metrics, 2006, pp. 1–6.

M. Claypool and J. Tanner, “The effects of jitter on the perceptual quality of video,” in Proc. ACM Int. Conf. Multimedia, 1999, pp. 115–118.

A. A. Khalek, C. Caramanis, and R. W. Heath, “Delay-constrained video transmission: Quality-driven resource allocation and scheduling,” IEEE J. Sel. Topics Signal Process., vol. 9, no. 1, pp. 60–75, Feb. 2015.

D. Li, T. Jiang, and M. Jiang, “Quality assessment of in-the-wild videos,” in Proc. ACM Int. Conf. Multimedia, 2019, pp. 2351–2359.

D. Amodet al., “Deep speech 2: End-to-end speech recognition in english and mandarin,” in Proc. Int. Conf. Mach. Learn., 2016, pp. 173–182.

K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

D. Fleet and Y. Weiss, “Optical flow estimation,” in Handbook of Mathematical Models in Computer Vision, Berlin, Germany: Springer, 2006, pp. 237–257.

C. Bailer, B. Taetz, and D. Stricker, “Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 4015–4023.

E. Wood, T. Baltrusaitis, C. Hewitt, S. Dziedzio, T. J. Cashman, and J. Shotton, “Fake it till you make it: Face analysis in the wild using synthetic data alone,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 3681–3691.

J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual Losses for Real-Time Style Transfer and Super-Resolution,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 694–711.

M. Cooke, J. Barker, S. Cunningham, and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition,” J. Acoustical Soc. Amer., vol. 120, no. 5, pp. 2421–2424, 2006.

N. D. Narvekar and L. J. Karam, “A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection,” in Proc. IEEE Int. Workshop Qual. Multimdia Experience, 2009, pp. 87–91.

J. S. Chung and A. Zisserman, “Out of time: Automated lip sync in the wild,” in Proc. Asian Conf. Comput. Vis. Workshop Multi-View Lip-Reading, 2016, pp. 251–263.

A. Bulat and G. Tzimiropoulos, “How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks),” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 1021–1030.

T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, “Learning a model of facial shape and expression from 4D scans,” ACM Trans. Graph., vol. 36, no. 6, pp. 194:1–194:17, 2017.

H. Li, S. Wang, P. He, and A. Rocha, “Face anti-spoofing with deep neural network distillation,” IEEE J. Sel. Topics Signal Process., vol. 14, no. 5, pp. 933–946, Aug. 2020.

Jun Ling received the B.E. degree in electronic engineering and information science from the University of Science and Technology of China, Hefei, China, in 2018. He is currently working toward the Ph.D. degree with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include image animation, talking face generation, and deep generative modeling.

Xu Tan (Senior Member, IEEE) received the master’s degree from Zhejiang University, Hangzhou, China, in 2015. He is currently a Principal Research Manager with Microsoft Research Asia, Beijing, China. His research interests include deep learning and distributed machine learning, and their applications for NLP, machine translation, ranking in search, and recommendation.

Liang Chen received the B.E. degree from the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China, in 2018. He is currently working toward the Ph.D. degree in computer science and technology with Tsinghua University, Shenzhen, China. His research interests include deep learning, speech signal processing, text-to-speech synthesis, and talking face generation.

Runnan Li received the Ph.D. degree in computer science from Tsinghua University, Shenzhen, China, in 2019. He is currently a Postdoctoral Fellow and Senior Software Engineer with Cloud + AI Group, Microsoft, Beijing, China. He is currently working on empowering people to create with AI-generated content, including audio and video. His research interests include speech processing and generation, visual perception, and video generation.

Yuchao Zhang received the B.S. and M.S degrees in computer science and technology from Northwestern Polytechnical University, Xi’an, China, in 2016 and 2019, respectively. He is currently a Speech Engineer with Microsoft Cloud + AI Group, Beijing, China. His research interests include AI-generated content, such as speech synthesis, audio generation, and visemes.

Sheng Zhao (Senior Member, IEEE) shipped state of art text to speech into products with billions of users including Azure Cognitive Speech Service, Windows Speech Platform and Microsoft Edge Read Aloud, multimedia system. He was the Guest Editor of IEEE Transactions on Multimedia in 2018 and the Guest Editor of IEEE Transactions on Broadcasting, special issue on quality of experience for advanced broadcast services in 2018.

Li Song (Senior Member, IEEE) received the B.E. and M.S. degrees in engineering in 1997 and 2000, respectively, and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2005. He joined SJTU as a Faculty Member, where he is currently a Full Professor with the Department of Electronic Engineering. From 2011 to 2012, he was a Visiting Professor with Santa Clara University, Santa Clara, CA, USA. He has more than 200 publications, 40 granted patents, and 18 standard technical proposals in the field of video coding and image processing. His research interests include video processing and multimedia system. He was the recipient of the National Science and Technology Progress Award in 2015, Okawa Foundation Research Grant in 2012, Second Place Award from IEEE ICME-Twitch Grand Challenge in 2017, Best 10% Paper Award from the IEEE VCIP in 2016, and Best Paper Award from the International Conference on Wireless Communications and Signal Processing in 2010. He is a Technical Committee Member of Visual Signal Processing and Communications in the IEEE Circuits and Systems Society and was the area or session Chair for various international conferences and workshops. He has been an Associate Editor for Multidimensional Systems and Signal Processing since 2012 and the Guest Editor of IEEE TRANSACTIONS ON BROADCASTING, special issue on quality of experience for advanced broadcast services in 2018.