MODA: Mapping-Once Audio-driven Portrait Animation with Dual Attentions

Yunfei Liu[1] Lijian Lin[1] Fei Yu[2] Changyin Zhou[2] Yu Li[1]*
[1]International Digital Economy Academy (IDEA) [2]Vistring Inc.
https://tinyurl.com/iccv23-moda

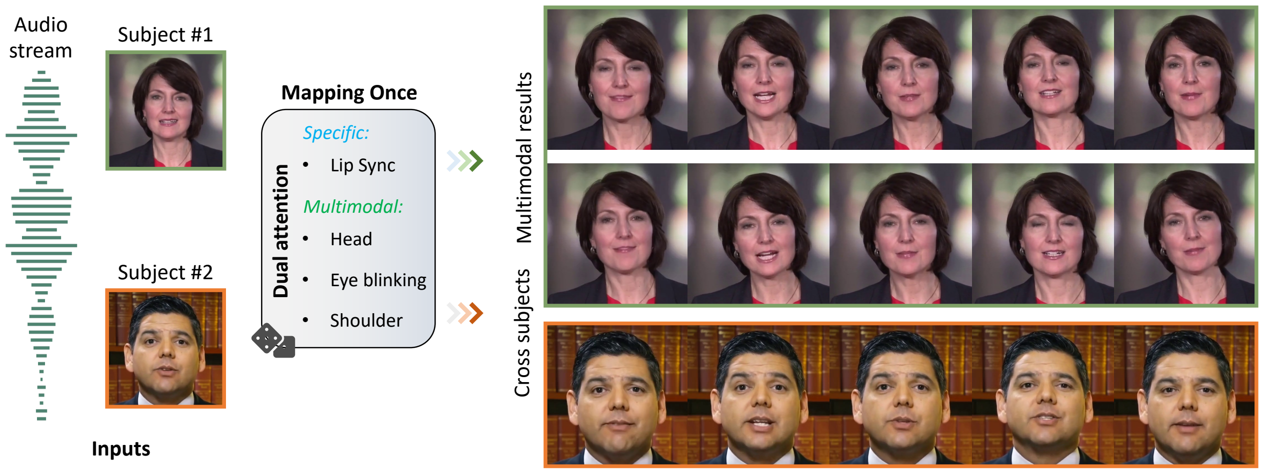

Figure 1: We propose a mapping-once system with dual-attention for multimodal and high-fidelity portrait video animation.

Abstract

*Audio-driven portrait animation aims to synthesize portrait videos that are conditioned by given audio. Animating high-fidelity and multimodal video portraits has a variety of applications. Previous methods have attempted to capture different motion modes and generate high-fidelity portrait videos by training different models or sampling signals from given videos. However, lacking correlation learning between lip-sync and other movements (*e.g.*, head pose/eye blinking) usually leads to unnatural results. In this paper, we propose a unified system for multi-person, diverse, and high-fidelity talking portrait generation. Our method contains three stages,* i.e.*, 1) Mapping-Once network with Dual Attentions (MODA) generates talking representation from given audio. In MODA, we design a dual-attention module to encode accurate mouth movements and diverse modalities. 2) Facial composer network generates dense and detailed face landmarks, and 3) temporal-guided renderer syntheses stable videos. Extensive evaluations demonstrate that the proposed system produces more natural and realistic video portraits compared to previous methods.*

1. Introduction

Given an input audio, talking portrait animation is to synthesize video frames of a person whose poses and expressions are synchronized with the audio signal [2, 3, 4, 18]. This audio-driven portrait video generation task has gained increasing attention recently and has a wide range of applications in digital avatars, gaming, telepresence, virtual reality (VR), video production *etc*. Conventional portrait video generation consumes intensive labor and time during setting up the background, make-up, lighting, shooting, and editing. Moreover, a re-shot is always required when there exists new textual content. In contrast, audio-driven talking video generation is more convenient and attractive which only requires a new audio clip to render a new video.

Previous methods [7, 29, 52] try to learn the correspondence between audio and frames. However, these methods usually ignore the head pose as it is hard to separate head posture from facial movement. Many 3D face reconstruction algorithm-based and GAN-based [8] methods estimate intermediate representations, such as 3D face shapes [6, 50], 2D landmarks [22, 54], or face expression parameters [49], to assist the generation process. However, such sparse representations usually lost facial details, leading to over-smooth [44]. Recently, the neural radiance field

*Corresponding author.

(NeRF) [10, 44] has been widely applied in talking head generation for high-fidelity results. However, the implicit neural representation is hard to interpret and control. In addition, these methods are usually person-specific and require extra training or adaptation time for different persons.

Although quite a number of attempts and progresses have been made in recent years, it is still challenging to generate realistic and expressive talking videos. As humans are extremely sensitive to identifying the artifacts in the synthesized portrait videos, it sets a very high standard for this technique to become applicable. We summarize the following key points that affect human perceptions: 1) **Correctness**. The synthesized talking portrait video should be well synchronized with the driven audio. 2) **Visual quality**. The synthesized video should have high resolution and contain fine detail components. 3) **Diversity**. Besides the lip motion needing to be exactly matched to the audio content, the motion of other components like eye blinking and head movement are not deterministic. They should move naturally as a natural human does.

To achieve these goals, previous approaches either map the mouth landmarks and the head pose separately by learning different sub-networks [22, 50], or only model the mouth movement while the head pose is obtained from the existing video [29, 52]. However, lacking correlation learning between lip-sync and other movements usually leads to unnatural results. In this paper, we propose a **m**apping-**o**nce network with **d**ual **a**ttentions (MODA), which is a unified architecture to generate diverse representations for a talking portrait, simplifying the computational steps. In order to combine synchronization and diversity of the talking portrait generation, we carefully design a dual-attention module to learn deterministic mappings (*i.e.*, the accurate mouth movements driven by audio) and probabilistic sampling (*i.e.*, the diverse head pose/eye blinking from time-to-time), respectively. To summarize, our contributions can be listed as follows:

- We propose a talking portrait system that generates multimodal photorealistic portrait videos with accurate lip motion. Comprehensive evaluations demonstrate our system can achieve state-of-the-art performance.
- We propose a unified mapping-once with dual attention (MODA) network for generating portrait representation from subject conditions and arbitrary audio.
- We propose 3 technical points for taking portrait generation: 1) A transformer-based dual attention module for generating both specific and diverse representations. 2) A facial composer network to get accurate and detailed facial landmarks. 3) A temporally guided renderer to synthesize videos with both high quality and temporal stabilization.

2. Related Works

Audio-driven portrait animation. Talking heads and facial animation are research hot-spots in the computer vision community. Extensive approaches [5, 55] explore audio-driven mouth animation and audio-driven facial animation. We focus on animating a portrait in this work. Many methods [4, 5, 29, 31, 56] aim to find the correspondence between audio and frames. A large number of technologies (such as flow-learning [14, 40, 51], memory bank [25, 35], *etc*.) are explored for the correctness of talking head generation. However, these methods usually ignore the head pose, torso motion, and eye blinking, which are essential for a natural talking portrait generation. To generate diverse talking heads, recent methods [19, 52] propose to embed other modalities to control emotions or head pose. However, these methods usually require additional inputs.

Recently, neural radiance field (NeRF) [10] has been widely applied in 3D-related tasks as it can accurately reproduce complex scenes with implicit neural representation. Several works [10, 43, 21, 44] leverage NeRF to represent faces with audio features as conditions. Despite the high-quality results achieved, the motion of generated results is usually unnatural. Besides, the learning and inference processes are time-consuming. More recently, some diffusion-based methods [32, 34] are proposed to generate talking heads. However, their speed is limited by a large number of sampling steps in the diffusion process. Some methods are based on the 3D face reconstruction and GAN [15, 22, 45]. They estimate intermediate representations such as 2D landmarks [39, 48, 54], 3D face shapes [15, 36] or facial expression parameters [42, 50], to assist the generation process. Unfortunately, such sparse representation usually lost facial details. In this paper, we propose to learn dense facial landmarks and upper body points through a unified framework for talking portrait generation. The intermediate representation contains facial details and other movements, which can be interpreted and controlled easily.

Transformers in audio-driven tasks. Transformer [38] is a strong alternative to both RNN and CNN. Researchers find it works well in multimodal scenarios. We refer readers to the comprehensive survey [16] for further information. Some recent works adopt transformers to generate results from different modalities, such as audio-to-text, language translation, music-to-dance, *etc*. The most related work is FaceFormer [6], which is a speech-driven 3D facial animation approach. They proposed two types of bias for the transformer to better align audio and 3D face animation.

Vision-based facial reenactment. Video-based facial reenactment is another technique related to audio-driven animation [33, 31]. There are many works to reenact faces with different techniques, such as adversarial learning, few-shot learning, or even one-shot facial animation. They usually adopt pre-defined facial landmarks or in an unsupervised

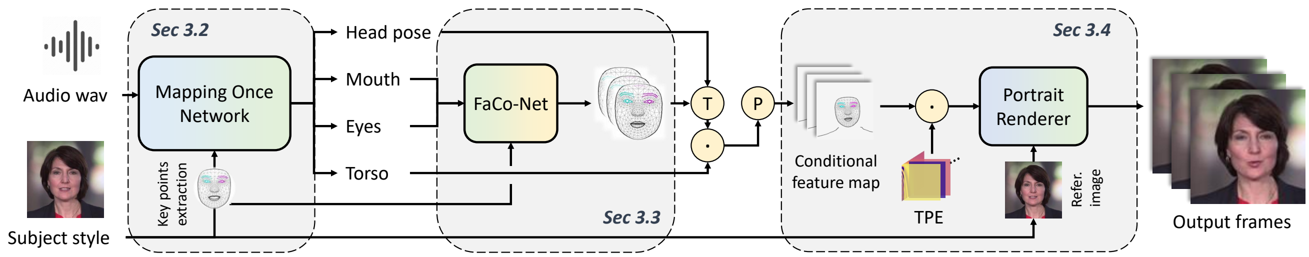

Figure 2: The proposed method is a three-stage system. Given the subject figure and arbitrary audio, the proposed system generates audio-driven video. Here Ⓣ denotes rigid transformation from canonical space to camera coordinate via head pose, $\odot$ denotes concatenation. Ⓟ is projection from 3D space to image coordinate. FaCo-Net is a facial composer network, which will be introduced in Sec. 3.3.

scheme. In another aspect, image-to-image translation (I2I) methods [20, 46] have also demonstrated impressive performance in converting images from one domain to another. However, single-frame renderers [20] ignore the temporal relations among video frames, leading to color jitters or unnatural background shakes in the final results. [41] proposes to use RNN [23] to capture the temporal relations among the input conditions, which generates stabilized results. However, these methods have difficulties in training [26]. In this paper, we find an alternative way to embed temporal information into I2I. Simply using temporal positional embedding [11] as an input condition, our method can achieve natural and stabilized results.

3. Methodology

We present a talking portrait system for high-fidelity portrait video generation with accurate lip motion and multi-modal motions, including head pose, eye blinking, and torso movements. The overall pipeline of this system is illustrated in Fig. 2. It contains three stages, 1) given the driven audio and conditioned subjects, mapping once talking portrait network with dual attentions (MODA) generates multimodal and correct semantic portrait components, 2) in the next, the facial composer network combines the facial components together and adds details for dense facial vertices, and 3) finally, a portrait renderer with temporally positional embedding (TPE) syntheses high-fidelity and stable videos.

3.1. Task Definition

In this section, we give the definition of the talking portrait task, which is to formulate a sequence-to-sequence translation manner [1] from talking portrait videos. Specifically, given a T-length audio sequence $\mathbf{A} = \{a_0, a_1, \dots, a_T\}$ with audio sampling rate r, a talking portrait method aims to map it into the corresponding video clip $\mathbf{V} = \{I_0, I_1, \dots, I_K\}$ with f frame-per-second (FPS), where $K = \lfloor fT/r \rfloor$. Since the data dimension of $\mathbf{V}$ is much larger than $\mathbf{A}$, many researchers propose to generate $\mathbf{V}$ progressively and introduce many types of intermediate representation $\mathbf{R}$. To make the generated $\mathbf{V}$ look natural, the constraint on $\mathbf{R}$ is critical. In previous audio-driven face animation approaches, $\mathbf{R}$ typically represents one type of face information, such as facial landmarks [22, 54] or head pose [50]. To better represent a talking portrait, we define $\mathbf{R}$ as the union of different portrait descriptors, *i.e.*, $\mathbf{R} = \{P^M, P^E, P^F, H, P^T\}$, where the elements of $\mathbf{R}$ are defined as follows,

1. Mouth points $P^M \in \mathbb{R}^{40\times 3}$. They have 40 points for representing mouth animation.
2. Eyes points $P^E \in \mathbb{R}^{60\times 3}$. They consist of eye and eyebrow points, which control eye blinking.
3. Facial points $P^F \in \mathbb{R}^{478\times 3}$. They contain dense facial 3D points for recording expression details.
4. Head pose $H \in \mathbb{R}^6$. It contains head rotations (θ, ϕ, ψ in Euler angle) and head transposes (x, y, z in Euclidean space).
5. Torso points $P^T \in \mathbb{R}^{18\times 3}$. They contain 18 points and each side of the shoulder is described by 9 points.

Note that P^M, P^E, and P^F are in canonical space for the convenience of face alignment. The process of talking portrait can be rewritten as $\mathbf{A} \rightarrow \mathbf{R} \rightarrow \mathbf{V}$. We design corresponding networks for these stages, respectively. The details are provided in the following subsections.

3.2. Mapping-Once Network with Dual Attentions

Mapping-once architecture. Given the driven audio $\mathbf{A}$ and subject condition $\mathbf{S}$, MODA aims to map them into $\mathbf{R}$ (consists of lip movement, eye blinking, head pose, and torso) with a single forward process. As illustrated in Fig. 3, the network in the first step contains three parts, *i.e.*, 1) two encoders for encoding audio features and extracting subject style, respectively, 2) a dual-attention module for generating diverse but accurate motion features, and 3) four tails for different motion synthesis. We first extract contextual features of the audio signal by Wav2Vec [30]. In the next, the extracted feature is projected into $s_a \in \mathbb{R}^{d\times T}$ via a multilayer perceptron (MLP), where d is the feature dimension for one frame and T denotes the number of frames of the

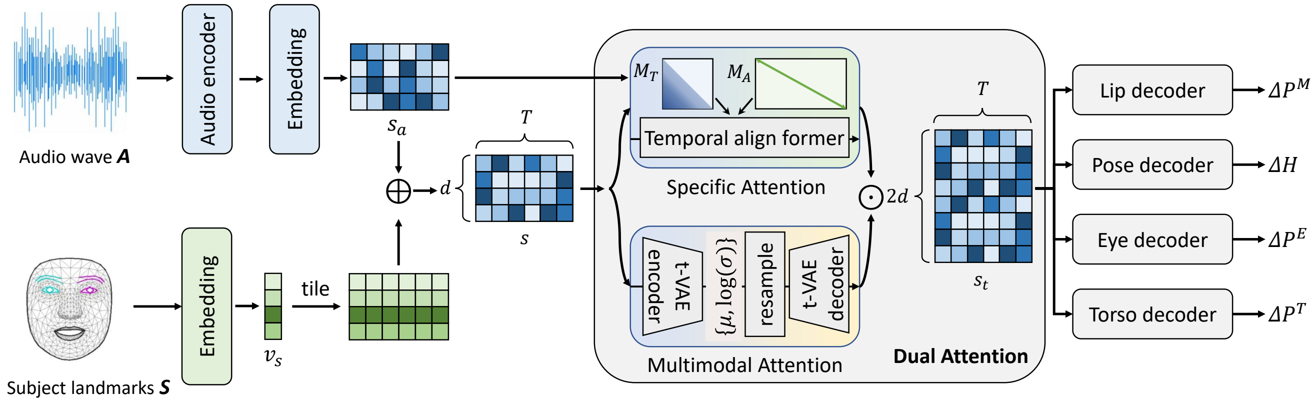

Figure 3: Architecture of MODA network. Given an audio and subject condition, MODA generates four types of motions within a single forward process. $\oplus$ denotes element-wise addition and $\odot$ is concatenation.

generated video. To model different speaker styles, we take the facial vertices of the conditioned subject as input. Then those vertices are projected to a d-dimensional vector v_s as the subject style code. Here the embedding layer is implemented by MLP. Next, s_a and v_s are combined as:

$$s = s_a \oplus \text{tile}(v_s), \tag{1}$$

where s is the combined feature, $\oplus$ is dimension-wise addition. Then the dual-attention module (DualAttn) takes s, s_a as input, and yields a temporally contextual version s_t,

$$s_t = \text{DualAttn}(s, s_a). \tag{2}$$

Next, we adopt 4 MLPs to decode the movements of lips P^M, head pose H, eye blinking P^E, and torso P^S, respectively. For each downstream task X, the computational process can be formulated as follows,

$$\Delta X = \Phi^X(s_t), \tag{3}$$

where $\Phi(\cdot)$ denotes an MLP and $\Delta X = X - \overline{X}$, $\overline{X}$ is extracted from referred subject image.

Dual-attention module. The talking portrait generation task is highly ill-posed since it requires generating multimodal results from limited-driven information. To solve this, we propose a dual-attention module that disentangles this task into a *specific mapping* and a *probabilistic mapping* problem. Specifically, this module generates 1) the temporally aligned feature for specific mapping between audio and lip movements, as well as 2) the temporally correlated feature for probabilistic mapping between audio and other movements of the talking portrait. To this end, we first design two sub-modules to learn these two different features, respectively. Then we fuse these two features via time-wise concatenation.

In detail, we propose a *specific attention* branch (SpecAttn) to capture the temporally aligned attention s_{sa} between s and audio feature s_a. Inspired by FaceFormer [6], our SpecAttn is formulated as:

$$\begin{aligned} s_{sa} &= \text{SpecAttn}(s_a, s) \\ &= \text{softmax}(\frac{\Gamma(s) \cdot s_a^T}{\sqrt{d}} + M_A)\Gamma(s), \end{aligned} \tag{4}$$

where d is the dimension of s_a, $\{\cdot\}^T$ indicates the transpose of the input parameter. The alignment bias $M_A(1 \leq i \leq T, 1 \leq j \leq T)$ is represented as:

$$M_A(i,j) = \begin{cases} 0, & i = j \\ -\infty, & \text{otherwise.} \end{cases} \tag{5}$$

Different from FaceFormer which performs cross-attention in an auto-regressive manner, we apply this operation on the entire sequence, which boosts the computation speed $T\times$ faster. In addition, to capture rich temporal information, we adopt a periodic positional encoding (PPE) and a biased casual self-attention on s (as in [6]):

$$s' = \Gamma(s) = \text{softmax}(\frac{\text{PPE}(s) \cdot \text{PPE}(s)^T}{\sqrt{d}} + M_T)\text{PPE}(s). \tag{6}$$

M_T is a matrix that has negative infinity in the upper triangle to avoid looking at future frames to make current predictions. M_T is defined as:

$$M_T(i,j) = \begin{cases} \lfloor (i-j)q \rfloor, & j \leq i, \\ -\infty, & \text{otherwise,} \end{cases} \tag{7}$$

where q is a hyper-parameter for tuning the sequence period. By doing this, the encoded feature s' contains rich spatial-temporal information, which aids the accurate talking portrait generation.

To generate vivid results and avoid the over-smoothing [44] representations, it is essential to learn the probabilistic mapping between the audio feature and portrait motions. We notice that Variational Autoencoder

(VAE) [17] can model probabilistic synthesis and shows many advanced performances in sequence generation tasks. Therefore, based on an advanced transformer Variational Autoencoder (t-VAE) [28], we design a *probabilistic attention* branch to generate diverse results. Formally, given the representation s, the probabilistic attention (ProbAttn) aims to generate a diverse feature s_{pa}. It first models the distribution of s with learned μ and σ through an encoder (Enc). Then it generates multimodal outputs through a re-sample operation with a decoder (Dec). The computational process is

$$\begin{aligned} \mu, \log\sigma &= \Phi^{\mu}(\text{Enc}(s)), \Phi^{\sigma}(\text{Enc}(s)), \\ s_{pa} &= \text{Dec}(x), \quad s.t.\ \ x \sim \mathcal{U}(\mu, \sigma), \end{aligned} \tag{8}$$

where Φ is an MLP. $\mathcal{U}(\mu, \sigma)$ is the Gaussian distribution with mean μ and variance σ. To force ProbAttn to learn diverse motion styles, we add Kullback–Leibler divergence (KLD) loss to constrain the feature from the bottleneck of t-VAE. The KLD loss is defined as follows:

$$\mathcal{L}_{KLD} = (-\frac{1}{2d_l}\sum^{d_l}(\log\sigma - \mu^2 - \sigma + 1), \tag{9}$$

where d_l is the dimension of μ. Finally, the dual-attention module outputs $s_t = s_{sa} \odot s_{pa}$ for downstream tasks.

Loss functions. The MODA has four decoders for generating talking portrait-related motions. To learn the mapping from the dual-attention module and four different types of motion, we adopt a multi-task learning scheme for MODA. Specifically, we minimize the L_1 distance between the ground-truth displacements and the predicted displacements. The loss can be written as

$$\begin{aligned} \mathcal{L}_{TP} = &\lambda_1|\Delta P^M_{gt} - \Delta P^M| + \lambda_2|\Delta R_{gt} - \Delta R| \\ &+ \lambda_3|\Delta P^E_{gt} - \Delta P^E| + \lambda_4|\Delta P^S_{gt} - \Delta P^S|, \end{aligned} \tag{10}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are hyper-parameters for balancing the different weights of downstream tasks. $|\cdot|$ is the L_1-norm. ΔP^*_{gt} and ΔP^* indicate the displacements of the ground truth and the predicted result, respectively. The total loss function is the sum of $\mathcal{L}_{TP}$ and $\mathcal{L}_{KLD}$, *i.e.*,

$$\mathcal{L}_{total} = \mathcal{L}_{TP} + \mathcal{L}_{KLD}. \tag{11}$$

3.3. Facial Composer Network

Given the subject information $\mathbf{S}$, the generated mouth points P^M, and eye points P^E, the facial composer network (FaCo-Net) aims to composite the refined facial dense landmarks. The generated facial dense landmarks $P^F =$ FaCo-Net$(\mathbf{S}, P^M, P^E)$. FaCo-Net consists of three encoders for consuming those three inputs and a decoder for facial landmarks generation. Similar to MODA, the subject encoder projects facial points $\mathbf{S}$ into a style code $\boldsymbol{p}_f$. The P^M and P^E are also projected to $\boldsymbol{p}_m$ and $\boldsymbol{p}_e$, which share the same latent space as $\boldsymbol{p}_f$. Next, $P^F = \Psi_c((\boldsymbol{p}_m \odot \boldsymbol{p}_e) \oplus \boldsymbol{p}_f)$, where Ψ_c is a facial dense point decoder. We adopt a vanilla GAN architecture [8] as the backbone of the discriminator (D). The FaCo-Net is trained to generate "realistic" facial dense points to fool D, whereas D is trained to distinguish the generated facial points from ground truths. The detailed architectures can be found in the supplementary materials. We use LSGAN loss [20] as the adversarial loss to optimize the D:

$$\mathcal{L}_{Disc}(D) = (z-1)^2 + \hat{z}^2, \tag{12}$$

where $z, \hat{z}$ is the discriminator output when inputting the ground-truth face points P^F_{gt} and the generated P^F, respectively. The loss for the generator is

$$\mathcal{L}_G = \mathcal{L}_{GAN}(\text{FaCo-Net}) + \lambda|P^F_{gt} - P^F|, \tag{13}$$

where P^F_{gt} is the ground-truth dense face landmarks. $\mathcal{L}_{GAN}(\text{FaCo-Net}) = (\hat{z}-1)^2$ is the adversarial loss, where $\hat{z} = D(P^F)$. The weight λ is empirically set to 10. After composition, the facial landmarks P^F are transformed to the camera coordinate via head pose H. The transformed facial landmarks and torso points P^T are projected into image space for photorealistic rendering.

3.4. Portrait Image Synthesis with TPE

The last stage of our system is a renderer that generates photorealistic facial renderings from previous predictions, as illustrated in Fig. 2. Specifically, we design a U-Net-like renderer G_R with TPE to generate both high-fidelity and stable videos. In our experiments, TPE is defined as

$$\begin{aligned} \text{TPE}_{(t,2i)} &= \sin(t * 2^i/100), \\ \text{TPE}_{(t,2i+1)} &= \cos(t * 2^i/100). \end{aligned} \tag{14}$$

$i = 0, 1, ..., 5$ is the dimension and t is the frame index. Then the rendered result t-frame I_t is generated with G_R:

$$I_t = G_R(I^c_t \odot I_r \odot \text{TPE}(t)), \tag{15}$$

where I^c_t is the condition image at frame index t. I_r is the reference image. The detailed architecture, training, and inference details are provided in our supplementary materials.

3.5. Implementation Details

Our models are trained on PyTorch [27] using Adam optimizer with hyper-parameters $(\beta_1, \beta_2) = (0.9, 0.99)$. The learning rate is set to 10^{-4} in all experiments.

We train all of our models on an NVIDIA 3090 GPU. It takes about (30, 2, 6) hours in total, (200, 300, 100) epochs with bath sizes of (32, 32, 4) for our three different stages, respectively. During testing, we select all the models with minimum validation loss. We use a sliding window (window size 300, stride 150) for arbitrary long input audio.

Table 1: Comparisons with state-of-the-art methods. † denotes our generated results with size 256×256 through a small renderer. The best results are highlighted in **bold**. The number with underline denotes the second-best result.

Method	Testset A from LSP [22]					Testset B from HDTF [51]				
	NIQE ↓	LMD-v ↓	LMD ↓	Sync ↑	MA ↑	NIQE ↓	LMD-v ↓	LMD ↓	Sync ↑	MA ↑
MakeItTalk (SIGGRAPH Asia'20 [54])	7.07	2.30	2.65	3.07	0.48	8.18	1.91	2.23	3.90	0.53
Wav2Lip (MM'20 [29])	7.31	1.95	1.81	**5.58**	0.64	7.83	2.08	1.97	**5.78**	0.51
Wav2Lip-GAN (MM'20 [29])	7.24	2.11	1.83	5.47	0.62	7.77	2.01	1.98	5.78	0.51
LSP (SIGGRAPH Asia'21 [22])	5.75	2.28	2.06	3.09	0.61	7.12	1.67	2.01	4.11	0.52
AD-NeRF (ICCV'21 [10])	5.81	2.89	2.77	2.98	0.41	-	-	-	-	-
SadTalker (CVPR'23 [50])	5.80	2.51	2.31	4.14	0.56	7.07	2.43	2.37	3.96	0.51
GeneFace (ICLR'23 [44])	6.61	2.22	2.17	3.08	0.65	-	-	-	-	-
Ground Truth (reference)	5.28	0.00	0.00	4.89	1.00	6.38	0.00	0.00	6.07	1.00
Ours †	5.77	**1.74**	1.51	4.52	0.70	7.05	1.60	2.04	4.34	**0.59**
Ours	**5.55**	1.79	**1.50**	4.48	**0.69**	**6.92**	**1.59**	**1.96**	4.16	0.56

4. Experiments

4.1. Experimental Setup

Dataset pre-processing. We evaluate our method on two publicly available datasets, *i.e.*, HDTF [51] and Video samples from LSP [22] (LSP dataset). Each video contains a high-resolution portrait with an audio track. The average video length is 1-5 minutes and we process them at 25 fps. We randomly select 80% of them for training and the remaining videos for evaluation. Specifically, we get 132 videos for training and 32 videos for evaluation. Each video is cropped to keep the face at the center and then resized to 512×512. The LSP dataset contains 5 different target sequences of 4 different subjects for training and testing. These sequences span a range of 3-5 minutes. All videos are extracted at 25 fps and the synchronized audio is sampled at $16K$ Hz frequency. We split videos as 80% / 20% for training and validation.

We detect 478 3D facial landmarks for all videos using Mediapipe*. Then we estimate the head pose H for all videos using method [9]. According to these head poses, the 3D facial landmarks are projected to the canonical space through rigid transformation. We extract the 3D mouth points, and eye-related points as P^M and P^E for each frame. The torso points are estimated from the boundary of the shoulders, which is detected through the face parsing algorithm†. For more data pre-processing details please refer to our supplement materials.

Evaluation metrics. We demonstrate the superiority of our method on multiple metrics that are widely involved in the talking portrait field. To evaluate the correctness of generated mouth, we use mouth landmark distance (LMD) and velocity of mouth landmark distance (LMD-v) between generated video and reference video in canonical space. In addition, we also calculate the Insertion-over-Union (IoU) for the overlap between the predicted mouth area and the ground truth area (MA). We use the confidence score from SyncNet (Sync) [29] to measure the audio-video synchronization. Since the result cannot be perfectly aligned with the ground-truth video, we use Natural Image Quality Evaluator (NIQE) [24] as the metric for image quality. NIQE is able to capture the naturalness of image details, it is widely used in blind image quality assessment.

Figure 4: Visual comparison of 5 methods.

*https://google.github.io/mediapipe/

†https://github.com/zllrunning/face-parsing.PyTorch

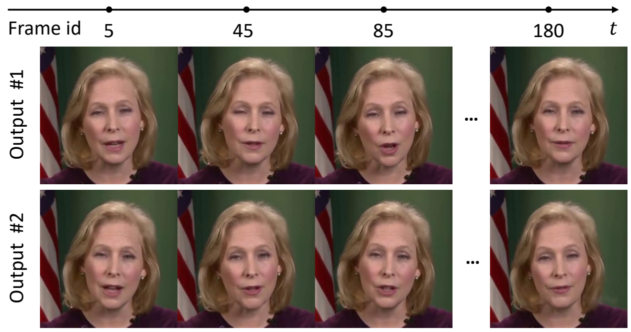

Figure 5: Multimodal results with the same mouth shape.

4.2. Quantitative Comparison

We compare our method with several state-of-the-art one-shot talking portrait generation works (LSP [22], MakeItTalk [54], Wav2Lip [29], AD-NeRF [10], SadTalker [50], and GeneFace [44]). For MakeItTalk, Wav2Lip, and SadTalker, the evaluation is performed on their publicly available checkpoint directly. Since these methods only generate low-resolution results, we retrained a small portrait renderer to generate low-resolution results for a fair comparison. The rest methods are retrained on our dataset under the same condition. Note that AD-NeRF and GeneFace are NeRF-based methods that are extremely time-consuming on all videos, we only provide the numerical results on the LSP dataset. As shown in Tab. 1, the proposed method achieves the best overall video quality (lowest NIQE, 5.25) and the correctness of audio-lip synchronization (lowest LMD, LMD-v, and highest MA). Our method also shows comparable performance with other fully talking-head generation methods in terms of lip-sync score. Please note that a higher sync score is not always lead to better results since it is too sensitive to the audio where unnatural lip movements may get a better score [50].

4.3. Qualitative Evaluation

User Study. We conduct user studies with 20 attendees on 30 videos generated by ours and the other methods. The driving audio is selected from four different languages: English, Chinese, Japanese, and German. The videos are generated across 5 subjects. Each participant is asked to select the best generated talking-portrait videos based on three major aspects: lip synchronization accuracy, the naturalness of movements including head movement, eye blinking, and upper body movement, and the video quality of the generated portrait. We collect the voting results and calculate the best-voting percentage of each method. The statistics are reported in Tab. 2. Overall, users prefer our results on lip synchronization, the naturalness of portrait, and video quality, indicating the effectiveness of the proposed method.

Qualitative comparison. Fig. 4 demonstrates the visual comparison among different methods. The results from LSP [22] have some warping effects without 3D consistency. Wav2Lip [29] can generate accurate mouth motions. However, their mouth areas usually have blurry boundaries and artifacts, which make the video unnatural. The results from AD-NeRF [10] have blurry boundaries of shoulders. SadTalker [50] may suffer from out of sync. GeneFace [44] has obvious artifacts on the neck region. Compared to these methods, our system generates portrait videos with overall high-quality and natural mouth movements.

Diverse outputs. Fig. 5 shows the diverse rendered videos that are driven by the same audio. These videos have different head poses, eye-blinking, and upper bodies while sharing the same mouth structures. These results demonstrate that our MODA network is able to generate vivid and diverse talking portrait videos.

4.4. Ablation Study and Performance Analysis

We conduct ablation studies on dual-attention in MODA, FaCo-Net, and TPE in portrait renderer.

Dual-attention module. We choose to 1) replace DualAttn with a multi-layer LSTM block [12]; 2) remove the specific attention branch and 3) remove the multimodal attention branch to evaluate the effectiveness of the dual-attention module. Numerical results on LSP test set are reported in Tab. 3. Using LSTM block cannot generate multimodal results and the diverse score (here we use the variance of the generated facial landmarks) drops to 0. When removing the specific attention branch from the dual attention block, the MODA generates the over-smoothed lip movement, which may be out of lip synchronization and has large LMD and LMD-v errors.

FaCo-Net. The FaCo-Net aims to generate natural and consistent representations for our portrait renderer. We carry out an ablation study on it by removing this stage and directly replacing the eye landmarks and mouth landmarks with facial dense landmarks. Fig. 6a shows that condition images without FaCo-Net contain incorrect connections in the lip area and lose face details, leading to low SSIM ($0.871 \to 0.843$), PSNR ($24.77 \to 21.96$) and NIQE ($5.55 \to 6.71$) rendered images (as in Tab. 4). These results consistently prove the effectiveness of FaCo-Net.

Temporally positional encoding. We adopt the temporal consistency metric to measure to evaluate the frame-wise consistency (TCM [37]) of the generated videos. Specifically, the TCM is defined as

$$\mathrm{TCM} = \frac{1}{T}\sum_{t}^{T} \exp\left(-\frac{||O_t - \mathrm{warp}(O_{t-1})||^2}{||V_t - \mathrm{warp}(V_{t-1})||^2} - 1\right), \quad (16)$$

where O_t and V_t represent the t^{th} frame in the referenced video (O) and generated video (V), respectively. $\mathrm{warp}(\cdot)$ is the warping function using the optical flow [13]. The 2-norm of a matrix $||\cdot||$ is the sum of squares of its elements.

Table 2: User study analyses measured by best-voting percentage. Higher is better.

	Low resolution (256 × 256)				High resolution (512 × 512)			
Approach	MakeItTalk [54]	Wav2lip [29]	SadTalker [50]	Ours	LSP [22]	AD-NeRF [10]	GeneFace [44]	Ours
Lip-sync accuracy	15.2%	30.5%	16.5%	37.6%	24.6%	7.9%	19.0%	48.5%
Naturalness of movement	12.8%	14.0%	18.6%	54.5%	19.0%	6.3%	7.1%	67.6%
Image quality	8.3%	7.2%	14.3%	70.0%	22.8%	11.1%	16.7%	49.7%

(a) Effectiveness of FaCo-Net.

(b) Effectiveness of TPE.

Figure 6: Ablation studies on FaCo-Net (a) and temporal positional encoding (b).

Table 3: Ablation study on MODA. Removing dual attention or replacing it with LSTM block has negative effects.

Method	LMD-v	LMD	Diverse
replace with LSTM	2.49	2.79	0
w/o multimodal attention	3.01	2.81	0
w/o specific attention	1.80	1.55	1.70
Final	**1.79**	**1.50**	1.57

Table 4: Ablation study on FaCo-Net.

Method	SSIM ↑	PSNR ↑	NIQE ↓
w/o FaCo-Net Net	0.843	21.96	6.71
Final	**0.871**	**24.77**	**5.55**

Table 5: Ablation study on TPE. Higher is better.

Method	TCM ↑
Renderer w/o TPE	0.63
Renderer with TPE	**0.71**

Through this equation, the generated video (V) is encouraged to be temporally consistent according to variations in the reference video (O). Fig. 6b demonstrates the comparison of video sequences with/without TPE. We find TPE can stabilize video synthesis, especially when training videos with changing backgrounds. Numerical results in Tab. 5 also show that TPE can increase TCM score.

5. Discussions and Conclusions

We present a deep learning approach for synthesizing multimodal photorealistic talking-portrait animation from audio streams. Our method can render multiple personalized talking styles with arbitrary audio. Our system contains three stages, *i.e.*, MODA, FaCo-Net, and a high-fidelity portrait renderer with temporal guidance. The first stage generates lip motion, head motion, eye blinking, and torso motion with a unified network. This network adopts a dual-attention mechanism and is able to generate diverse talking-portrait representations with correct lip synchronization. The second stage generates fine-grained facial dense landmarks powered by generated lip motion and eye blinking. Finally, we generate the intermediate representations for our temporal-guided renderer to synthesize both high-fidelity and stable talk-portrait videos. Experimental results and user studies show the superiority of our method. Analytical experiments have also verified different parts of our system.

Limitations and future work. While our approach achieves impressive results in a wide variety of scenarios, there still exist several limitations. Similar to most deep learning-based methods, our method cannot generalize well on unseen subjects or extremely out-of-domain audio. It may require fine-tuning the renderer for new avatars. We also looking forward to future work to find a person-invariant renderer to achieve high-quality synthesis without additional finetuning.

References

[1] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 3

[2] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *ECCV*, 2020. 1

[3] Lele Chen, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *CVPR*, 2019. 1

[4] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *CVPR*, 2017. 1, 2

[5] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: an animator-centric viseme model for expressive lip synchronization. *TOG*, 35(4):1–11, 2016. 2

[6] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *CVPR*, 2022. 1, 2, 4

[7] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *CVPR*, 2021. 1

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1, 5

[9] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *ECCV*, 2020. 6

[10] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *ICCV*, 2021. 2, 6, 7, 8, 12, 14

[11] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *PAMI*, 45(1):87–110, 2022. 3

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 7

[13] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 7

[14] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *SIGGRAPH*, 2022. 2

[15] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *TOG*, 36(4):1–12, 2017. 2

[16] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021. 2

[17] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 5

[18] Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and CV Jawahar. Towards automatic face-to-face translation. In *ACMMM*, 2019. 1

[19] Avisek Lahiri, Vivek Kwatra, Christian Frueh, John Lewis, and Chris Bregler. Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization. In *CVPR*, 2021. 2

[20] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NeurIPS*, 2017. 3, 5

[21] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-aware implicit neural audio-driven video portrait generation. In *ECCV*. Springer, 2022. 2

[22] Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live speech portraits: real-time photorealistic talking-head animation. *TOG*, 40(6):1–17, 2021. 1, 2, 3, 6, 7, 8, 11, 12, 14

[23] Larry R Medsker and LC Jain. Recurrent neural networks. *Design and Applications*, 5:64–67, 2001. 3

[24] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 6

[25] Se Jin Park, Minsu Kim, Joanna Hong, Jeongsoo Choi, and Yong Man Ro. Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory. In *AAAI*, 2022. 2

[26] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *ICML*, 2013. 3

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019. 5

[28] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. *arXiv preprint arXiv:2104.05670*, 2021. 5

[29] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACMMM*, 2020. 1, 2, 6, 7, 8, 12

[30] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. Wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019. 3

[31] Oliver Schreer, Roman Englert, Peter Eisert, and Ralf Tanger. Real-time vision and speech driven avatars for multimedia applications. *TMM*, 10(3):352–360, 2008. 2

[32] Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized talking head synthesis. *arXiv preprint arXiv:2301.03786*, 2023. 2

[33] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *NeurIPS*, 32, 2019. 2

[34] Michał Stypułkowski, Konstantinos Vougioukas, Sen He, Maciej Zięba, Stavros Petridis, and Maja Pantic. Diffused heads: Diffusion models beat gans on talking-face generation. *arXiv preprint arXiv:2301.03396*, 2023. 2

[35] Anni Tang, Tianyu He, Xu Tan, Jun Ling, Runnan Li, Sheng Zhao, Li Song, and Jiang Bian. Memories are one-to-many mapping alleviators in talking face generation. *arXiv preprint arXiv:2212.05005*, 2022. 2

[36] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *ECCV*, 2020. 2

[37] S. Varghese, Y. Bayzidi, A. Bar, N. Kapoor, and T. Fingscheidt. Unsupervised temporal consistency metric for video segmentation in highly-automated driving. In *CVPR Workshop*, 2020. 7, 11

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[39] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020. 2

[40] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. In *IJCAI*, 2021. 2

[41] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *NeurIPS*, 2018. 3

[42] Xinsheng Wang, Qicong Xie, Jihua Zhu, Lei Xie, et al. Anyonenet: Synchronized speech and talking head generation for arbitrary person. *arXiv preprint arXiv:2108.04325*, 2021. 2

[43] Shunyu Yao, RuiZhe Zhong, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering. *arXiv preprint arXiv:2201.00791*, 2022. 2

[44] Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, JinZheng He, and Zhou Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. In *ICLR*, 2023. 1, 2, 4, 6, 7, 8, 14

[45] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137*, 2020. 2

[46] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017. 3

[47] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *IJCV*, 129:3051–3068, 2021. 12

[48] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *ICCV*, 2019. 2

[49] Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. Facial: Synthesizing dynamic talking face with implicit attribute learning. In *ICCV*, 2021. 1

[50] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *CVPR*, 2023. 1, 2, 3, 6, 7, 8

[51] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *CVPR*, 2021. 2, 6

[52] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *CVPR*, 2021. 1, 2

[53] Yijun Zhou and James Gregson. Whenet: Real-time fine-grained estimation for wide range head pose. *arXiv preprint arXiv:2005.10353*, 2020. 11

[54] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *TOG*, 39(6):1–15, 2020. 1, 2, 3, 6, 7, 8, 12

[55] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemenet: Audio-driven animator-centric speech animation. *TOG*, 37(4):1–10, 2018. 2

[56] Hao Zhu, Huaibo Huang, Yi Li, Aihua Zheng, and Ran He. Arbitrary talking face generation via attentional audio-visual coherence learning. In *IJCAI*, 2020. 2