Hierarchically Controlled Deformable 3D Gaussians for Talking Head Synthesis

Zhenhua Wu[1,2], Linxuan Jiang[3], Xiang Li[4], Chaowei Fang[5], Yipeng Qin[6], Guanbin Li[1,7,8*]

[1]Sun Yat-sen University,
[2]Shanghai Innovation Institute,
[3]Guangdong University of Technology,
[4]Gezhi Intelligent Technology,
[5]Xidian University,
[6]Cardiff University,
[7]Peng Cheng Laboratory
[8]Guangdong Key Laboratory of Big Data Analysis and Processing
wuzhh56@mail2.sysu.edu.cn, 2112105077@mail2.gdut.edu.cn, lixiang@tideo.cn, chaoweifang@outlook.com,
qiny16@cardiff.ac.uk, liguanbin@mail.sysu.edu.cn

Abstract

Audio-driven talking head synthesis is a critical task in digital human modeling. While recent advances using diffusion models and Neural Radiance Fields (NeRF) have improved visual quality, they often require substantial computational resources, limiting practical deployment. We present a novel framework for audio-driven talking head synthesis, namely *Hierarchically Controlled Deformable 3D Gaussians (HiCoDe)*, which achieves state-of-the-art performance with significantly reduced computational costs. Our key contribution is a *hierarchical control* strategy that effectively bridges the gap between sparse audio features and dense 3D Gaussian point clouds. Specifically, this strategy comprises two control levels: i) coarse-level control based on a 3D Morphable Model (3DMM) and ii) fine-level control using facial landmarks. Extensive experiments on the HDTF dataset and additional test sets demonstrate that our method outperforms existing approaches in visual quality, facial landmark accuracy, and audio-visual synchronization while being more computationally efficient in both training and inference.

Introduction

Audio-driven talking head synthesis, a classical task in digital human modeling, is widely used in digital broadcasting, virtual reality (VR), and teleconferences. This cross-modal task aims to modify facial expressions and lip movements in the target video to match a given input audio clip while preserving facial continuity and fine details like teeth structure.

Existing solutions for audio-driven talking head synthesis can be broadly classified into two categories: 2D-based and 3D-based methods. Recent state-of-the-art 2D approaches predominantly leverage advanced deep generative models, particularly diffusion models (Rombach et al. 2022). While these methods yield impressive visual fidelity, they often require substantial computational resources and suffer from slow rendering speeds, limiting their practical applicability. Similarly, cutting-edge 3D-based methods frequently employ Neural Radiance Fields (NeRF) (Mildenhall et al. 2021), which are also computationally intensive. For instance, training a NeRF-based talking head model can take more than 8 hours on high-end GPUs (Li et al. 2023). This poses significant challenges for deploying NeRF-based methods in real-world applications. Therefore, the development of a realistic and efficient audio-driven talking head synthesis solution remains an open challenge.

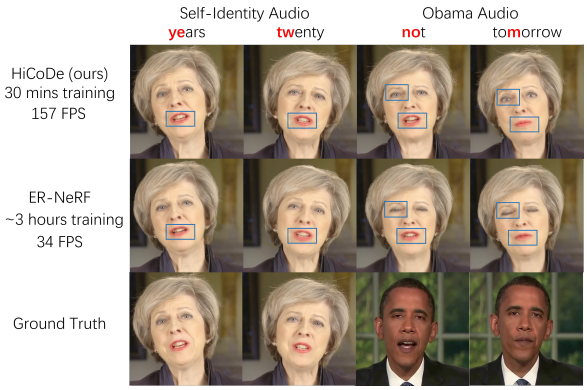

Figure 1: Our HiCoDe method significantly outperforms the state-of-the-art NeRF-based method, ER-NeRF (Li et al. 2023), in fidelity (highlighted with blue boxes), as well as computational costs (about five times faster in training and inference). Our model allows talking head synthesis using both self-identity audio and other identity audio. Please zoom in for more details.

In this paper, we address this challenge by proposing a novel framework based on Deformable 3D Gaussians (Yang et al. 2023), namely HiCoDe, which can significantly reduce training and inference time while achieving better rendering quality. In a nutshell, our framework consists of four basic components: i) audio encoding, ii) deformable 3D Gaussians, iii) rendering, and iv) head-background composition, which is a novel combination of off-the-shelf methods. However, a naive combination of these components proves ineffective due to an inherent mismatch between the *sparse*

*Corresponding author is Guanbin Li.

audio features and the *dense* 3D Gaussian point clouds. This mismatch significantly impairs the efficacy of audio-driven facial expression control, resulting in suboptimal performance. To fix this mismatch, we draw inspiration from a key insight that although 3D Gaussians are *dense*, the head they represent is constrained to a specific identity and the facial expressions are *sparse*. Thus, we propose to decouple head identity from facial expressions and use the input *sparse* audio to only control similarly *sparse* facial expressions. Accordingly, we propose a novel *hierarchical control* strategy that effectively bridges input audio with 3D Gaussians: at the coarse level, we utilize 3D Morphable Model (3DMM) (Blanz and Vetter 2003) parameters as a bridge, capitalizing on their ability to generate smooth facial animations; at the fine level, we incorporate facial landmarks for precise local adjustments. This hierarchical strategy is motivated by the complementary strengths of these two representations: 3DMM offers robust global control but suffers from limited expressiveness due to its constrained set of base models, while facial landmarks provide high-precision local control but lack comprehensive structural information. By synergizing these complementary attributes, our method achieves both coherent global animation and nuanced local expressiveness. Extensive experimental results on HDTF (Zhang et al. 2021) and two testing sets (Li et al. 2023; Ye et al. 2023b) demonstrate that our method achieves state-of-the-art performance in terms of visual quality, facial landmark accuracy, and visual-audio synchronization, and is more computationally efficient in training and inference. We acknowledge a concurrent work, GaussianTalker (Cho et al. 2024), which also explores the application of deformable 3D Gaussians for audio-driven talking head synthesis. While this approach represents a significant step in the field, it relies on a straightforward attention mechanism to modulate 3DGS attributes using audio features. This is suboptimal as it uses *sparse* audio features to control the *dense* 3D Gaussian point clouds, producing artifacts in output videos. In contrast, our proposed hierarchical control strategy addresses these limitations, offering more stable and visually coherent results.

In summary, our main contributions are as follows:

- We propose a novel framework for audio-driven talking head synthesis based on deformable 3D Gaussians, namely HiCoDe, which generates realistic and computationally-efficient facial animations.
- We propose a novel *hierarchical control* strategy that effectively bridges input audio with 3D Gaussians, comprising three components: i) a landmark-based attention mechanism, ii) a landmark-based fine-grained control strategy, iii) and a 3DMM-based coarse control strategy.
- Extensive experiments on the HDTF and other test sets demonstrate that our method greatly outperforms existing ones in both visual quality and computational cost.

Related Work

2D-Based Talking Portrait Synthesis. Early 2D-based approaches (Prajwal et al. 2020; Jamaludin, Chung, and Zisserman 2019; Chen et al. 2019) relied on image-based constraints to model mouth shapes. For example, some methods (Xie et al. 2021; Zhou et al. 2020; Suwajanakorn, Seitz, and Kemelmacher-Shlizerman 2017) utilized facial landmarks to control facial key points, enabling audio-driven facial animation. However, these approaches often introduce additional errors and result in blurry and distorted generated outputs. Recently, there are several 2D methods based on diffusion models (Shen et al. 2023; Stypułkowski et al. 2024; Yu et al. 2023). However, they require significant computational resources and are slow in rendering, limiting their practical applicability.

3D-Based Talking Portrait Synthesis. There are two major types of 3D-based methods. Between them, the first type relies on 3D parametric facial models that capture the geometric and textural information of the face, such as 3D Morphable Models (3DMM) (Blanz and Vetter 2003) and FLAME (Li et al. 2017). For example, Lu, Chai, and Cao (2021) proposed a method where audio features are directly mapped to the facial parameters of a model to render and synthesize the corresponding person's face. Song et al. (2022) proposed a method to enhance the expressiveness of the facial animation parameters by decoupling identity information from the parameters. However, due to the limitations of their rendering approaches, these methods struggle to generate detailed facial animations. The second type is based on Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) and has been very successful. For example, Radnerf (Tang et al. 2022) decomposes the human body into two parts, namely the head and the torso, and reconstructs them using separate NeRF models. The audio input is then connected through a neural network. ERNerf (Li et al. 2023) improves the control method by incorporating attention mechanisms and enhances rendering details using triplanes. GeneFace (Ye et al. 2023b) addresses Radnerf's average face issue by incorporating three-dimensional facial landmark information. GeneFace++ (Ye et al. 2023a) further addresses the problem of mapping one audio input to multiple facial expressions. However, these methods are computationally heavy and also struggle to generate detailed facial animations because aligning multiple mappings is very challenging. To address these issues, we propose incorporating 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) as the backbone for 3D-based talking portrait synthesis. Concurrent with our work, GaussianTalker (Cho et al. 2024) also uses 3DGS but employs a simple attention module to connect audio features and 3DGS attributes, resulting in jitter and gaps between the generated face and background. Therefore, controlling the deformation of 3DGS point clouds with audio features remains a challenge.

3D Gaussian Splatting. It is a novel explicit 3D reconstruction approach (Kerbl et al. 2023) that is significantly faster to train than NeRF. Specifically, it uses a 3D Gaussian representation to reconstruct the content of objects and achieves high rendering speeds through splatting-based rendering, which has been applied to a variety of 3D reconstruction tasks.

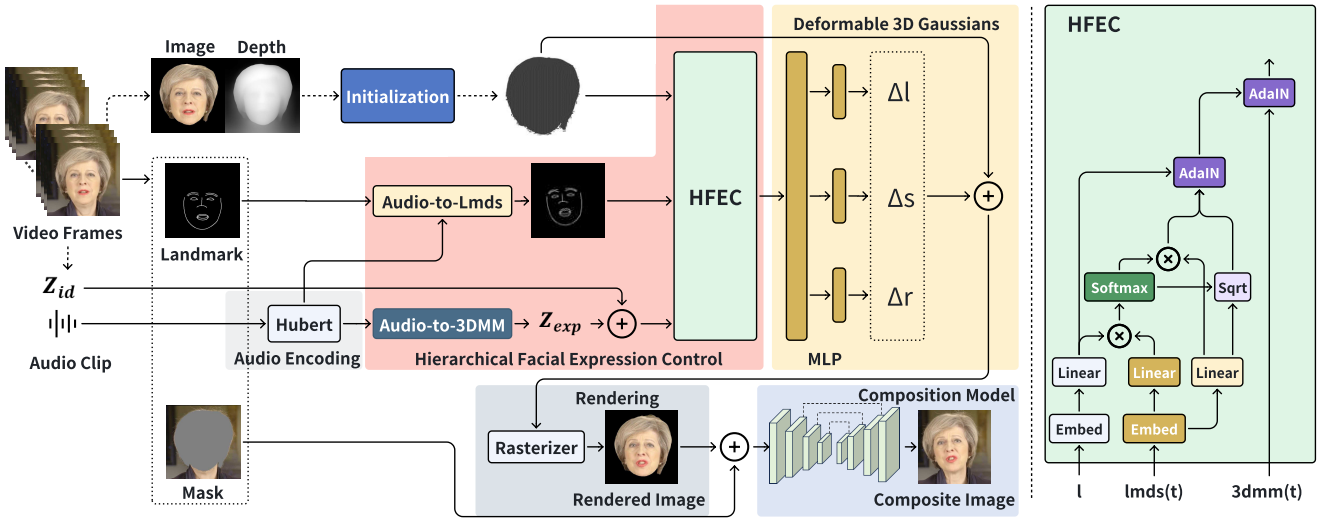

Figure 2: Overview. Left: Our framework consists of four basic components: i) audio encoding, ii) deformable 3D Gaussians, iii) rendering, and iv) head-background composition, which is a novel combination of off-the-shelf methods. However, making such a combination work is non-trivial. Addressing this challenge, we propose a novel *Hierarchical Facial Expression Control* (HFEC) strategy that effectively bridges input audio with 3D Gaussians: at the coarse level, we utilize 3D Morphable Model (3DMM) (Blanz and Vetter 2003) parameters as a bridge, capitalizing on their ability to generate smooth facial animations; at the fine level, we incorporate facial landmarks for precise local adjustments. Right: Detailed structure of our HFEC module.

Preliminaries

3D Gaussians. Following the 3DGS framework, we use a set of N 3D Gaussian spheres gs^i ($i \in \{1, 2, 3, ..., N\}$) to represent the scene at each frame, achieving real-time photorealistic rendering via differentiable rasterization:

$$gs^i = (l^i, r^i, s^i, o^i, c^i) \tag{1}$$

where $l^i \in \mathbb{R}^3$ denotes location, $r^i \in \mathbb{R}^4$ denotes rotation, $s^i \in \mathbb{R}^3$ denotes scale, and $o^i \in \mathbb{R}$ denotes opacity and $c^i \in \mathbb{R}^3$ denotes color. We represent the learned gs^i after training as its *canonical form*:

$$gs^i_{\text{cn}} = (l^i_{\text{cn}}, r^i_{\text{cn}}, s^i_{\text{cn}}, o^i_{\text{cn}}, c^i_{\text{cn}}) \tag{2}$$

3D Gaussian Rendering. According to (Zwicker et al. 2001), the resulting 3D Gaussians can be projected onto a 2D image plane I, where differential rendering is applied to each pixel using the following 2D covariance matrix Σ':

$$\Sigma' = JV\Sigma V^T J^T \tag{3}$$

where J is the Jacobian matrix approximating the affine projection transformation; V denotes the view matrix that transforms points from world coordinates to camera coordinates; Σ represents the original covariance matrix of the 3D Gaussians. To facilitate the learning of 3D Gaussians, Σ is further decomposed using a rotation matrix R and an anisotropic scaling matrix S:

$$\Sigma = RSS^T R^T \tag{4}$$

where R and S are constructed with the rotation r^i and scale s^i of each 3D Gaussian sphere gs^i, respectively. During rendering, we stack the m closest Gaussians at each pixel of according to depth order, forming the composite color for that pixel as a weighted blend of contributing Gaussians:

$$c_{\text{pixel}} = \sum_{j=1}^{m} \omega^j c^j = \sum_{j=1}^{m} \alpha^j \prod_{k=1}^{j-1} (1 - \alpha^k) c^j \tag{5}$$

$$\alpha^j = \text{sigmoid}(o^j) e^{-\frac{1}{2}(l_{\text{pixel}} - l^j)^T \Sigma' (l_{\text{pixel}} - l^j)} \tag{6}$$

where l_{pixel} is the location of the pixel on 2D image plane, and α denotes the view-dependent opacity.

Method

Following previous works (Fan et al. 2022), we define *audio-driven talking head synthesis* as modifying the facial expressions and lip movements in a given target video to match a given input audio clip. In this work, we propose to solve this task with a novel framework based on Deformable 3D Gaussians (Yang et al. 2023), featuring a *hierarchical facial expression control* strategy based on facial landmarks and the 3D Morphable Model (3DMM).

Framework Overview

As shown in Fig. 2, our framework consists of four basic components: i) audio encoding; ii) deformable 3D Gaussians; iii) rendering; and iv) composition of the generated head and background. We will briefly introduce them below and then detail our *hierarchical facial expression control* strategy in the following subsections.

Audio Encoding. Following (Fan et al. 2022), we feed the input audio to a pre-trained self-supervised speech model, HuBERT (Hsu et al. 2021), to extract robust acoustic features from unlabeled speech corpora at each timestamp t.

Deformable 3D Gaussians. To achieve a more refined and faster generation of controllable human faces, we draw inspiration from (Yang et al. 2023) and decouple facial motion from model geometry. Specifically, we represent the geometry of the 3D face model with a point cloud and learn a deformation network F_Δ to predict its motion, i.e., the offset between the point cloud attributes (r^i, s^i, l^i) at time t and those of its learned canonical form:

$$F_\Delta(l^i_{\text{cn}}, 3\text{dmm}(t), \text{lmds}(t)) = (\Delta r^i_t, \Delta s^i_t, \Delta l^i_t) \quad (7)$$

where l^i_{cn} denotes the location of gs^i in its learned canonical form, $3\text{dmm}(t)$ denotes the control signal produced by our novel audio-to-3DMM module, and $\text{lmds}(t)$ denotes the control signal produced by our novel audio-to-landmark module. Then, we have:

$$r^i_t = r^i_{\text{cn}} + \Delta r^i_t, s^i_t = s^i_{\text{cn}} + \Delta s^i_t, l^i_t = l^i_{\text{cn}} + \Delta l^i_t \quad (8)$$

Rendering. We render the 3D Gaussians at time t using the same process described in Preliminaries, where (r^i_t, s^i_t, l^i_t) are obtained from Eq. 8, and $o^i_t = o^i_{\text{cn}}, c^i_t = c^i_{\text{cn}}$ are obtained from their learned canonical form.

Head-background Composition. Given rendered 3D Gaussians representing a reenacted head at time t, we employ a pix2pix-inspired (Isola et al. 2017) approach to seamlessly composite the rendered image onto the corresponding input video frame, synthesizing the final output frame of the talking head animation. Empirically, we observed that our approach effectively addressed the discontinuities and ghosting artifacts produced by end-to-end reenactment models at composition boundaries.

Remark on 3DGS Initialization. A well-chosen initialization for the point cloud can provide the model with more prior knowledge, aiding in the optimization process. To better leverage facial prior knowledge, we propose a novel 3DGS initialization strategy that leverages monocular depth estimation to incorporate effective facial priors. Specifically, we first randomly select a frame from the input video and estimate its depth map using Depth-anything (Yang et al. 2024). Then, we employ (Deng et al. 2019) to extract the head region R_{head} from the selected frame. Utilizing this information, we construct a dense 3D point cloud where each point corresponds to a pixel in R_{head}. The (x, y) coordinates of each point are derived from the 2D pixel coordinates in the image plane, while the z coordinate is obtained from the corresponding depth value in the estimated depth map. Accordingly, we initialize the point locations l with (x, y, z). We initialize the other attributes r, s, o, c using the same methods in the native 3DGS (Kerbl et al. 2023).

Hierarchical Facial Expression Control

The main challenge in incorporating 3DGS into audio-driven talking head synthesis is to effectively control the facial expressions represented by 3D point clouds based on the input audio cues, which is a key component of the deformation network F_Δ (Eq. 7). This is particularly demanding due to the inherent mismatch between the *sparse* audio features and the *dense* 3DGS point clouds. We address this by observing that although 3DGS point clouds are *dense*, the head they represent is of the same identity for each video and the facial expressions are *sparse*. Thus, we propose to decouple head identity from facial expressions and use the input *sparse* audio to only control similarly *sparse* facial expressions, which is consistent with the decoupling of geometry and motion in deformable 3D Gaussians. In our work, we implement this with the 3D Morphable Model (3DMM) (Blanz and Vetter 2003) parameterized with z_{exp} and z_{id} that separately represent expression and identity with a set of bases.

[Coarse] Audio-to-3DMM Module. We feed the acoustic features extracted by the Audio Encoding (HuBERT) component of our framework into a Faceformer network (Fan et al. 2022) to learn the mapping between input audio and the expression parameters $z_{\text{exp}} \in \mathbb{R}^{64}$ of the 3DMM model. Following (Paysan et al. 2009), we obtain its corresponding identity parameter $z_{\text{id}} \in \mathbb{R}^{80}$ from the input video, which remains static across frames. We denote the output of this module as $3\text{dmm}(t)$ at time t.

However, while 3DMM is effective, its expressive power is limited by the relatively small number of its base models (a.k.a., **coarse**), making it difficult to fit the fine-grained details of facial animation. To address this limitation, we propose a novel fine-grained audio-to-landmark module that greatly enhances the details of output facial animations.

[Fine] Audio-to-Landmark Module. Unlike the 3DMM that relies on base models, we first randomly select 20 frames from the input video and then employ the Google MediaPipe (Lugaresi et al. 2019) to extract 478 facial keypoints from each frame. The facial keypoints of all 20 frames are regraded as the identity information to generate audio-driven facial landmarks. Specifically, we feed these 478×20 landmarks and input audio into a Faceformer network (Fan et al. 2022) and generate the corresponding output landmarks for each target frame respectively. We denote the output of this module as $\text{lmds}(t)$ at time t.

To maximize the efficacy of both the coarse 3DMM and fine-grained landmark control signals in guiding the deformation of 3D Gaussians, we introduce a novel hierarchical control module as follows.

Hierarchical Facial Expression Control Module. The key insight driving this module is the intrinsic nature of 3D Gaussian primitives, defined by their **mean** and **covariance** matrices. We leverage this characteristic by employing Adaptive Instance Normalization (AdaIN) layers (Huang and Belongie 2017) to modulate these parameters, enabling precise control over the spatial distribution and shape of each Gaussian primitive. Accordingly, as Fig. 2 shows, this module consists of three components:

- **Landmark-based Attention Mechanism.** Leveraging the correspondence between landmarks (i.e., facial keypoints) and 3D Gaussian points, we enforce the landmark constraints by introducing a novel landmark-based attention mechanism whose mean attention matrix W_μ is computed as follows:

$$W_\mu = V \times \text{softmax}(Q^T \times K)^T \quad (9)$$

where Q denotes the embeddings of 3D Gaussians' loca-

tions l, K and V are the embeddings of $\mathrm{lmds}(t)$, and $\times$ represents the matrix product. Similarly, we compute the variance attention matrix W_σ as follows:

$$W_\sigma = \sqrt{(V \circ V) \times \mathrm{softmax}(Q^T \times K)^T - (W_\mu \circ W_\mu)}, \tag{10}$$

where $\circ$ denotes the Hadamard product.

- **Landmark-based Fine-grained Control.** Then, we use (W_μ, W_σ) to modulate Q, which is the embedding of input 3D Gaussians l through an AdaIN layer:

$$y = \mathrm{AdaIN}(l, (W_\mu, W_\sigma)) = W_\sigma(\frac{Q - \mu(Q)}{\sigma(Q)}) + W_\mu \tag{11}$$

- **3DMM-based Coarse Control.** Denote the mean and standard deviation of $3\mathrm{dmm}(t)$ as Z_μ and Z_σ, respectively. We use them to modulate y through the other AdaIN layer:

$$M = \mathrm{AdaIN}(y, (Z_\mu, Z_\sigma)) = Z_\sigma(\frac{y - \mu(y)}{\sigma(y)}) + Z_\mu \tag{12}$$

Remark on the Choice of AdaIN. Empirically, we justify our choice of AdaIN to modulate 3D Gaussians by comparing it to alternative strategies, including simple concatenation and attention-based methods. All these alternative methods failed to produce satisfactory results.

Model Training

The main components of our framework can be formulated as independent sub-tasks and trained separately:

Audio-to-3DMM Module. This module is trained in a supervised manner. To create the dataset, we first split a set of given videos into paired audio and video streams; then, we use a pretrained 3DMM model to extract the 3DMM parameters from each video frame; finally, we align the extracted 3DMM parameter sequences with their corresponding audio clips. During training, the network is optimized using a mean squared error (MSE) loss function:

$$L_{\text{3DMM}} = \sum_{t=1}^{T} ||3\mathrm{dmm}_{gt}(t) - 3\mathrm{dmm}(t)||^2 \tag{13}$$

where t is the timestamp, and $3\mathrm{dmm}_{gt}(t)$ represents the ground truth of 3DMM parameters.

Audio-to-Landmark Module. Similar to the training of the audio-to-3DMM module, we create a paired audio and landmarks dataset using Mediapipe (Lugaresi et al. 2019) to obtain ground truth landmarks $\mathrm{lmds}_{gt}(t)$. We also use a mean squared error (MSE) loss function for its training:

$$L_{\text{lmds}} = \sum_{t=1}^{T} ||\mathrm{lmds}_{gt}(t) - \mathrm{lmds}(t)||^2 \tag{14}$$

Expression-Controlled 3DGS Head Synthesis. We extract the face regions from video frames and formulate the training of this component as a *face image reconstruction* task, which consists of three stages:

- *Initialization Stage (3,000 iterations)*: only the first video frame is used. The deformation network is not activated. Point adding and pruning operations are allowed.
- *Deformation Stage (2,000 iterations)*: all video frames are used. The deformation network is activated. Point adding and pruning operations are allowed.
- *Refinement Stage (15,000 iterations)*: all video frames are used. The deformation network is activated. Point adding and pruning operations are not allowed.

In all three stages, we employ the same loss function to optimize both the deformation network and 3D Gaussians:

$$L_{\text{3DGS}} = \lambda_1 L_{\text{RGB}} + \lambda_2 L_{\text{SSIM}} + \lambda_3 L_{\text{VGG}} + \lambda_4 L_{\text{WING}} \tag{15}$$

where

$$\begin{aligned} L_{\text{RGB}} &= ||I^f - I^f_{gt}|| + ||I^m - I^m_{gt}|| \\ L_{\text{SSIM}} &= ||\text{SSIM}(I^f) - \text{SSIM}(I^f_{gt})|| \\ L_{\text{VGG}} &= ||F_{\text{VGG}}(I^f) - F_{\text{VGG}}(I^f_{gt})|| \\ &\quad + ||F_{\text{VGG}}(I^m) - F_{\text{VGG}}(I^m_{gt})|| \\ L_{\text{WING}} &= \frac{1}{N}\text{Wing}(lmds, lmds_{gt}) \end{aligned} \tag{16}$$

where I^f and I^f_{gt} are the rendered face image and its corresponding ground truth, respectively; I^m is the mouth region of I^f and I^m_{gt} is its corresponding ground truth; $F_{\text{VGG}}(\cdot)$ calculates the features from the first four layers of a pre-trained VGG network (Sengupta et al. 2019); l and l_{gt} are the landmarks of I^f and I^f_{gt} extracted by MediaPipe (Lugaresi et al. 2019), respectively; $\text{Wing}(\cdot)$ is the Wing loss (Feng et al. 2018) measuring the difference between two sets of landmarks. Note that we have included additional loss terms for the mouth regions to enhance their rendering quality.

Head-background Composition. We formulate the training of this module as a supervised image composition task whose inputs are paired face and background images and the output is their composited image. To effectively train it, we employ a hybrid loss function comprising reconstruction and GAN terms (Abdal, Qin, and Wonka 2019):

$$L_{\text{composition}} = L_{\text{recon}} + L_{\text{GAN}} \tag{17}$$

where

$$\begin{aligned} L_{\text{recon}} &= \lambda_4 ||I - I_{gt}||^2 + \lambda_5 ||F_{\text{VGG}}(I) - F_{\text{VGG}}(I_{gt})|| \\ L_{\text{GAN}} &= \sum_i \frac{1}{n(D_i)} ||D_i(I) - D_i(I_{gt})||_1 \\ &\quad + \log D(I_{gt}) + \log(1 - D(I)) \end{aligned} \tag{18}$$

where I and I_{gt} denote the composited image and its corresponding ground truth; D_i denotes the i-th layer in discriminator D and $n(D_i)$ is the number of elements in it; L_{GAN} comprises a feature matching term and an adversarial term.

Experiments

Experimental Setup

Datasets. We aim to synthesize high-fidelity talking face images with various audio inputs, which necessitates the use

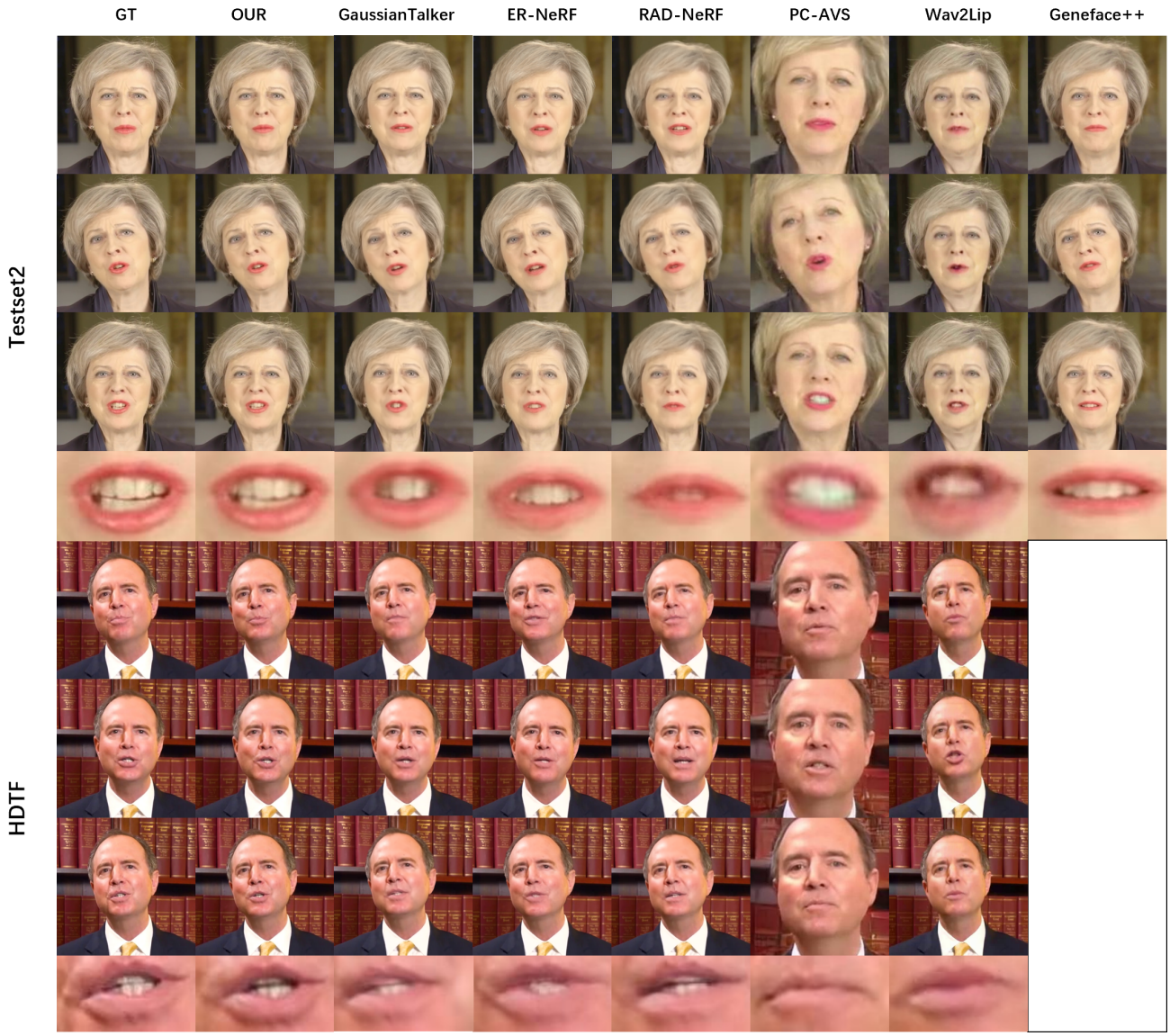

Figure 3: Qualitative comparisons with SOTA methods on one ID from HDTF and Testset2. Geneface++ is excluded from HDTF due to the unavailability of a pre-trained model. The fourth and last rows depict the mouth details of the row above, where our method maintains consistent mouth shapes with the ground truth, and the teeth and lip details are noticeably superior.

of high-resolution datasets with dimensions of 512×512 or greater. Following (Li et al. 2023; Ye et al. 2023b,a; Tang et al. 2022), we conduct experiments on three datasets: HDTF (Zhang et al. 2021), Testset 1 (Li et al. 2023), and Testset 2 (Ye et al. 2023b). For the HDTF dataset, we randomly select 8 videos (corresponding to 8 distinct subjects) with four videos of female subjects and four videos of male subjects to ensure gender balance.

Data Preprocessing. All videos used in our experiments are extracted at 25 frames per second (FPS), and the corresponding audio waveforms are sampled at 16 kHz. Subsequently, we decompose the videos into individual frames and perform the following preprocessing steps: i) Crop out the subject's head from each frame; ii) Resize the cropped frames to a resolution of 512×512; iii) Segment the head region from the background and extract relevant masks. For each video, we employ a 3D Morphable Model (3DMM) to extract identity, expression, and pose parameters for every frame. We also use mediapipe(Lugaresi et al. 2019) to extract the landmarks. We obtain the target person's identity parameters by averaging the identity parameters across all frames in the video for the same individual. We use Depth-Anything (Yang et al. 2024) to estimate the depth of the first frame for each video ID.

Baselines and Evaluation Metrics. We choose two categories of baselines for comparison: i) 2D-based methods including Wav2Lips (Prajwal et al. 2020) and PC-AVS (Zhou et al. 2021); ii) 3D-based methods, including

Method	HDTF					Testset 1					Testset 2					Training	Inference
	SSIM↑	PSNR↑	LPIPS↓	LMD↓	AVConf↑	SSIM↑	PSNR↑	LPIPS↓	LMD↓	AVConf↑	SSIM↑	PSNR↑	LPIPS↓	LMD↓	AVConf↑	Time	FPS
Ground Truth	1	N/A	0	0	8.960	1	N/A	0	0	8.610	1	N/A	0	0	8.442	-	-
Wav2Lip	0.759	20.354	0.138	4.452	**8.451**	0.659	15.950	0.289	5.309	7.809	0.647	19.507	0.217	3.544	**7.943**	-	15
PC-AVS	0.756	22.483	0.154	4.926	8.694	0.787	21.579	0.140	6.678	7.544	0.757	22.557	0.156	4.771	8.809	-	31
ER-NeRF	0.932	30.214	0.035	2.914	6.936	0.965	35.221	0.018	2.619	5.708	0.819	28.297	0.060	2.850	6.309	3h	34
RAD-NeRF	0.939	29.911	0.063	2.976	6.480	0.959	33.849	0.039	2.824	6.644	0.838	28.715	0.133	3.207	6.443	8h	38
Geneface++	-	-	-	-	-	-	-	-	-	-	0.825	28.763	0.119	3.019	7.229	22h	24
GaussianTalker	0.964	32.622	0.024	2.074	7.507	0.973	36.637	0.019	2.211	6.389	0.962	33.195	0.021	2.095	6.579	2h	98
Ours	**0.983**	**37.767**	**0.013**	**1.819**	8.428	**0.983**	**38.107**	**0.010**	**1.528**	**7.961**	**0.985**	**39.892**	**0.014**	**1.216**	7.861	**30min**	**157**

Table 1: Quantitative comparisons with state-of-the-art methods, including Wav2Lip (Prajwal et al. 2020), PC-AVS (Zhou et al. 2021), ER-NeRF (Li et al. 2023), RAD-NeRF (Tang et al. 2022), and GaussianTalker (Cho et al. 2024)). We evaluated Testset2 using Geneface++'s pre-trained model, as its training code is unavailable. **Bold**: best results; Underline: second-best results.

RADNerf (Tang et al. 2022), ERNerf (Li et al. 2023), Geneface++ (Ye et al. 2023a) and GaussianTalker (Huang et al. 2023). We divide the experiments into two parts: i) audio inputs are from the same identity of training data; ii) audio inputs are from identities different to training data. In the experiments, we use the following metrics for evaluation: i) PSNR, SSIM (Wang et al. 2004), and LPIPS (Zhang et al. 2018) that measure the visual quality of the generated images; ii) LMD (Chen et al. 2018) and Audio-Visual Confidence (AVConf) (Chung and Zisserman 2017) that evaluate the accuracy of face expression.

Quantitative Comparison

As shown in Table 1, our method outperforms state-of-the-art ones on most metrics. The AVConf of our method is marginally lower than that of Wav2Lip as Wav2Lip only modifies the mouth region, which brings synchronization advantages but leads to image blurring and pose fixation. In contrast, our method modifies the entire facial region, achieving a better balance between visual quality and synchronization accuracy. For example, on HDTF, our method achieves 0.019 higher SSIM and 0.155 lower LMD than the second-best method GaussianTalker, demonstrating the superiority of our method in image synthesis quality and facial structure prediction accuracy. Also, our method demonstrates significantly reduced training time and improves rendering efficiency compared to NeRF-based methods.

Qualitative Comparison

Compared to existing methods, our approach excels in capturing facial details and achieving synchronization with audio input. Specifically, as Fig. 3 shows, i) PC-AVS loses a significant amount of individual identity information and produces blurry facial images. ii) Wav2Lip cannot synchronize with facial poses; iii) ER-NeRF and RAD-NeRF synchronize poorly with the input audio and produce artifacts, especially in the teeth region, and blurry lips. They also lose hair details at the boundaries of the head region. iv) Geneface++ also synchronizes poorly with the input audio. v) GaussianTalker, a concurrent work to our method, also produces blurry mouths, synchronizes poorly with the input audio, and loses hair details at the boundaries of the head region. In contrast, our method demonstrates excellent performance in capturing mouth details, synchronizes well with the input audio, and preserves fine-grained details of the input video (e.g., hair at the boundaries of the head region). Similar conclusions also hold for audio inputs from different identities (Fig. 1).

Method	SSIM↑	PSNR↑	LPIPS↓	LMD↓	AVConf↑
w/o 3DMM	0.970	35.042	0.023	2.187	5.706
w/o Landmarks	0.979	37.564	0.018	1.868	7.460
None	0.969	32.614	0.027	5.281	3.597
w/ Addition	0.970	34.749	0.021	2.418	3.847
w/ Attention	0.968	33.511	0.026	2.966	2.934
w/ Random Init.	0.974	37.141	0.013	1.998	7.386
w/o HFEC	0.957	31.738	0.037	4.258	3.476
Ours (full)	**0.984**	**38.285**	**0.011**	1.511	8.403

Table 2: Ablation study on i) components of our *hierarchical facial expression control* module, including "None" (rows 2, 3 and 4); ii) our choice of AdaIN against *Addition* and *Attention* as alternative strategies (rows 5 and 6); iii) our 3D Gaussians initialization strategy against *Random Initialization* (row 7); iiii) our choice of using HFEC or not (row 8).

Ablation Study

First, we conduct an ablation study on our *hierarchical facial expression control* module. As shown in rows 2 and 3 of Table 2, both 3DMM and Landmark control signals contribute effectively to our final solution. "None" (row 4), i.e., controlling 3D Gaussians with audio features directly, works the worst, further demonstrating the effectiveness of our control strategy. Second, we conduct an ablation study on the choice of AdaIN in the implementation of our hierarchical control module. As shown in rows 5 and 6 of Table 2, alternative strategies including addition and attention fail to accurately control facial expressions (low LMD and AVConf), which justifies our choice of AdaIN. Third, we conduct an ablation study on our 3D Gaussians initialization strategy. As shown in row 7 of Table 2, our initialization significantly outperforms the original random initialization strategy. Finally we attempt to directly concatenate the two features and use them to control our model, which yields much worse results, reinforcing the necessity of our idea.

Conclusion

We present a novel framework for audio-driven talking head synthesis based on deformable 3D Gaussians, addressing the challenge of achieving high visual quality while maintaining computational efficiency. The key contribution of our work is the *hierarchical control* strategy, which bridges the gap between sparse audio inputs and dense 3D Gaussians and leverages the complementary power of 3DMM and facial landmark control signals. Our extensive experiments on the HDTF dataset and additional test sets demonstrate that our method outperforms existing state-of-the-art approaches in both visual quality and computational costs.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NO. 62322608, 62376206), in part by the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No.VRLAB2023A01), in part by the CCF-KuaiShou Fund(NO. 2024007), and in part by the CAAI-MindSpore Open Fund, developed on OpenI Community.

References

Abdal, R.; Qin, Y.; and Wonka, P. 2019. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF international conference on computer vision*, 4432–4441.

Blanz, V.; and Vetter, T. 2003. Face recognition based on fitting a 3D morphable model. *IEEE Transactions on pattern analysis and machine intelligence*, 25(9): 1063–1074.

Chen, L.; Li, Z.; Maddox, R. K.; Duan, Z.; and Xu, C. 2018. Lip movements generation at a glance. In *Proceedings of the European conference on computer vision (ECCV)*, 520–535.

Chen, L.; Maddox, R. K.; Duan, Z.; and Xu, C. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7832–7841.

Cho, K.; Lee, J.; Yoon, H.; Hong, Y.; Ko, J.; Ahn, S.; and Kim, S. 2024. GaussianTalker: Real-Time High-Fidelity Talking Head Synthesis with Audio-Driven 3D Gaussian Splatting. *arXiv preprint arXiv:2404.16012*.

Chung, J. S.; and Zisserman, A. 2017. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, 251–263. Springer.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

Fan, Y.; Lin, Z.; Saito, J.; Wang, W.; and Komura, T. 2022. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18770–18780.

Feng, Z.-H.; Kittler, J.; Awais, M.; Huber, P.; and Wu, X.-J. 2018. Wing loss for robust facial landmark localisation with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2235–2245.

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3451–3460.

Huang, R.; Lai, P.; Qin, Y.; and Li, G. 2023. Parametric implicit face representation for audio-driven facial reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12759–12768.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, 1501–1510.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Jamaludin, A.; Chung, J. S.; and Zisserman, A. 2019. You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision*, 127: 1767–1779.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4).

Li, J.; Zhang, J.; Bai, X.; Zhou, J.; and Gu, L. 2023. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7568–7578.

Li, T.; Bolkart, T.; Black, M. J.; Li, H.; and Romero, J. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.*, 36(6): 194–1.

Lu, Y.; Chai, J.; and Cao, X. 2021. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (TOG)*, 40(6): 1–17.

Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.-L.; Yong, M. G.; Lee, J.; et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Paysan, P.; Knothe, R.; Amberg, B.; Romdhani, S.; and Vetter, T. 2009. A 3D face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, 296–301. Ieee.

Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, 484–492.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Sengupta, A.; Ye, Y.; Wang, R.; Liu, C.; and Roy, K. 2019. Going deeper in spiking neural networks: VGG and residual architectures. *Frontiers in neuroscience*, 13: 95.

Shen, S.; Zhao, W.; Meng, Z.; Li, W.; Zhu, Z.; Zhou, J.; and Lu, J. 2023. DiffTalk: Crafting Diffusion Models for Generalized Audio-Driven Portraits Animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1982–1991.

Song, L.; Wu, W.; Qian, C.; He, R.; and Loy, C. C. 2022. Everybody's talkin': Let me talk as you want. *IEEE Transactions on Information Forensics and Security*, 17: 585–598.

Stypułkowski, M.; Vougioukas, K.; He, S.; Zięba, M.; Petridis, S.; and Pantic, M. 2024. Diffused heads: Diffusion models beat gans on talking-face generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5091–5100.

Suwajanakorn, S.; Seitz, S. M.; and Kemelmacher-Shlizerman, I. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4): 1–13.

Tang, J.; Wang, K.; Zhou, H.; Chen, X.; He, D.; Hu, T.; Liu, J.; Zeng, G.; and Wang, J. 2022. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. *arXiv preprint arXiv:2211.12368*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Xie, T.; Liao, L.; Bi, C.; Tang, B.; Yin, X.; Yang, J.; Wang, M.; Yao, J.; Zhang, Y.; and Ma, Z. 2021. Towards realistic visual dubbing with heterogeneous sources. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1739–1747.

Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891*.

Yang, Z.; Gao, X.; Zhou, W.; Jiao, S.; Zhang, Y.; and Jin, X. 2023. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. *arXiv preprint arXiv:2309.13101*.

Ye, Z.; He, J.; Jiang, Z.; Huang, R.; Huang, J.; Liu, J.; Ren, Y.; Yin, X.; Ma, Z.; and Zhao, Z. 2023a. GeneFace++: Generalized and Stable Real-Time Audio-Driven 3D Talking Face Generation. *arXiv preprint arXiv:2305.00787*.

Ye, Z.; Jiang, Z.; Ren, Y.; Liu, J.; He, J.; and Zhao, Z. 2023b. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. *arXiv preprint arXiv:2301.13430*.

Yu, Z.; Yin, Z.; Zhou, D.; Wang, D.; Wong, F.; and Wang, B. 2023. Talking head generation with probabilistic audio-to-visual diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7645–7655.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, Z.; Li, L.; Ding, Y.; and Fan, C. 2021. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3661–3670.

Zhou, H.; Sun, Y.; Wu, W.; Loy, C. C.; Wang, X.; and Liu, Z. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4176–4186.

Zhou, Y.; Han, X.; Shechtman, E.; Echevarria, J.; Kalogerakis, E.; and Li, D. 2020. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6): 1–15.

Zwicker, M.; Pfister, H.; Van Baar, J.; and Gross, M. 2001. EWA volume splatting. In *Proceedings Visualization, 2001. VIS'01.*, 29–538. IEEE.