VividTalker: Generalized and Stable 3D Gaussian-based Talking Head Synthesis

Jie Pan, Ling Lei, Tianhang Tang, Shaobing Gao, and Yiguang Liu(✉)

College of Computer Science, Sichuan University, Chengdu, China

liuyg@scu.edu.cn

Abstract

Recently, 3D Gaussian Splatting (3DGS) has shown impressive performance in synthesizing realistic speech-driven portrait videos. However, previous 3DGS-based methods perform poorly in lip synchronization due to the limitation of small training data. Additionally, the synthesized portraits suffer from jittery poses due to inaccurate head pose parameter estimation. In this work, we propose VividTalker, a highly synchronized talking face generation method based on 3DGS, which produces stable pose movements and lip shapes corresponding to various audios. Specifically, our method consists of a Audio2Lip module and a Video2Pose module. In the Audio2Lip, we pretrain an Audio2Mesh module on a newly constructed large-scale 3D audiovisual dataset, and then perform fine-tuning on the target subject to adapt its speaking style. For the Video2Pose, we used a head feature point matcher and a body pose stabilizer to optimize head pose parameters, which were then fed into the head pose stabilizer to achieve natural head movements. We also introduced head-aware features to mitigate the head-torso separation issue. Extensive experiments demonstrate that our method can render lifelike talking videos with better visual quality compared to previous approaches.

Keywords: talking head synthesis, lip-synchronization, stable pose, 3D Gaussian Splatting

1. Introduction

Synthesized audio-driven talking head videos hold significant potential in various domains, including virtual reality [27], film production [15], and human-computer interaction [34]. Recent advancements in view synthesis techniques such as Neural Radiance Fields (NeRF) [22] and 3D Gaussian Splatting (3DGS) [14] have greatly propelled this field. These methods produce talking head videos with photorealistic visual quality, representing a substantial leap in fidelity over previous approaches.

However, compared to some traditional methods based on Generative Adversarial Networks (GAN) [35, 43, 45], although NeRF-based and 3DGS-based methods [9, 17, 31, 33] can maintain identity information of different portrait lip shapes, such as lip size and thickness, they fail to accurately simulate lip movements under different audio conditions. Additionally, there is an urgent issue of noticeable jittering in the synthesized portraits during speech, which is unacceptable for generating a realistic talking head.

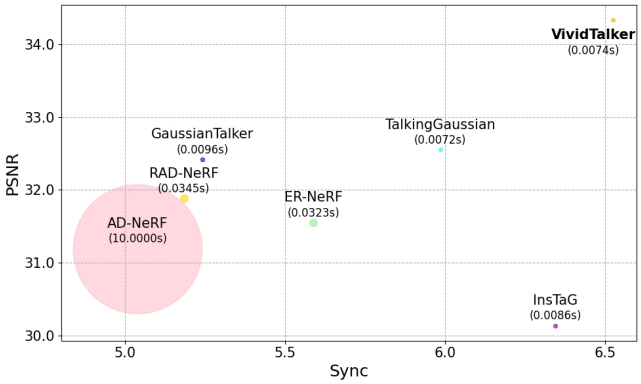

Fig. 1. Performance comparison between existing 3D-based talking face synthesis models [4, 9, 17, 18, 19, 33] and our approach in terms of fidelity, audio synchronization, and inference efficiency. Our method, VividTalker, delivers superior performance while maintaining inference speed comparable to TalkingGaussian. Bubble size indicates the per-frame inference time of each method.

To address aforementioned challenges, we propose VividTalker, a head talking synthesis framework utilizing 3D Gaussian Splatting technology to generate synchronized realistic talking portrait videos. Our approach is composed of two primary modules: the Audio2Lip, which enhances lip movement synchronization across various audio conditions, and the Video2Pose, which produces motion performances consistent with human perceptual expectations.

In the Audio2Lip, we introduce prior knowledge of lip movements through an Audio2Mesh module, which aids in alleviating the constraints of small sample learning. Specifically, we constructed a novel 3D audiovisual dataset comprising various speech audios and their corresponding lip

mesh sequences. Subsequently, we pretrained an Audio2Mesh module to predict the lip mesh for the given audio inputs and employed these meshes to drive the deformation of 3D Gaussian primitives at the lip positions. Due to the Audio2Mesh module's lack of consideration for the speaking styles of different portraits, a domain gap arises between the generated lip mesh and the target portrait's lip mesh. To address this issue, we propose a full fine-tuning process to align the predicted lip mesh with the target subject's distribution.

For Video2Pose module, departing from previous methods [17, 19] that depend on a limited number of facial keypoints to estimate head pose, we introduce a head feature point detector and matcher inspired by Simultaneous Localization And Mapping (SLAM) to gather more comprehensive supervisory information. By employing bundle adjustment to calculate projection loss, we achieve more precise estimations of head rotation angles and translation parameters. Additionally, we incorporate a head pose stabilizer to mitigate posture jitter that can arise from varying optimization levels across video frames. To generate a motion-stable torso, we implicitly learn the pose variation rule of the torso by conditioning on the head pose parameters obtained by the above method. Moreover, we integrate head perception features as additional input conditions to the model, mitigating the issue of head-torso separation.

The primary contributions of this paper are summarized as follows:

- We propose an Audio2Lip module that integrates prior information on lip movements from a 3D audiovisual dataset into the training of dynamic Gaussian radiance fields, resulting in more precise lip synchronization.
- We design a Video2Pose module that fully utilizes information between video frames to generate smooth and continuous pose movements. Furthermore, we mitigate the problem of head-torso separation.
- We construct a novel 3D audiovisual dataset, which includes audio from various speaking subjects and their corresponding lip mesh sequences. This dataset is available for research purposes.
- Comprehensive experiments indicate that the proposed VividTalker produces more realistic and synchronized talking portrait videos in terms of lip synchronization and visual quality.

2. Related Work

2.1. 2D-Based Talking Head Synthesis

Certain methods [28, 35, 43] ensure that audio matches lip movements by modifying facial regions, particularly the lips. For instance, Wav2Lip [28] introduces a lip synchronization expert to supervise the accuracy of lip movements. However, these methods typically use a few reference frames to reconstruct the lips, which hampers their ability to maintain the subject's identity. Moreover, some other approaches [12, 38, 41] aim to generate talking animations from a single image. For example, EAMM [12] and Real3D-Portrait [38] create facial animations of talking heads from just one image. Nevertheless, these methods often struggle to produce natural lip and body movements and fail to preserve the subject's identity, leading to unrealistic visual representations. In contrast, our VividTalker leverages 3D Gaussian Splatting (3DGS) [14] to synthesize talking portrait videos. By deforming 3DGS in canonical space, we can represent dynamic 3D scenes while preserving the consistency of the subject's identity and retaining scene details.

2.2. 3D-Based Talking Head Synthesis

Early 3D-based approaches often utilized 3D Morphable Models (3DMM) [13, 32] as intermediate representations to inject 3D prior knowledge. However, the coarse modeling of human heads by 3DMM led to additional errors. With the advent of Neural Radiance Fields (NeRF) [22], this technique has been employed to tackle the task of head talking synthesis. AD-NeRF [9] represents the first end-to-end method based on NeRF. Building on AD-NeRF, RAD-NeRF [33] integrates Instant-NGP [25], thereby accelerating both training and inference speeds. ER-NeRF [19] achieves a compact and efficient rendering approach by introducing a tri-plane hash encoder and a regional attention mechanism. GeneFace [37] and SyncTalk [26] improve the realism of head talking synthesis videos through the use of pre-trained audiovisual encoders. Although the above methods based on NeRF have achieved acceptable performance, their inference speed remains a limitation. Some 3DGS-based methods [4, 17, 18] addresses facial distortion issues in NeRF-based methods and achieves efficient video synthesis, yet it still encounters challenges with lip-audio synchronization and posture jitter. In contrast, our proposed VividTalker excels with its advantages in audio-lip synchronization, natural body posture, and real-time inference, enhancing both the realism and efficiency of video synthesis.

3. Method

In this section, we present the proposed method, VividTalker. Our approach consists of two primary components: the Audio2Lip and the Video2Pose, as depicted in Fig. 2. The detailed descriptions of these components are provided in the subsequent subsections.

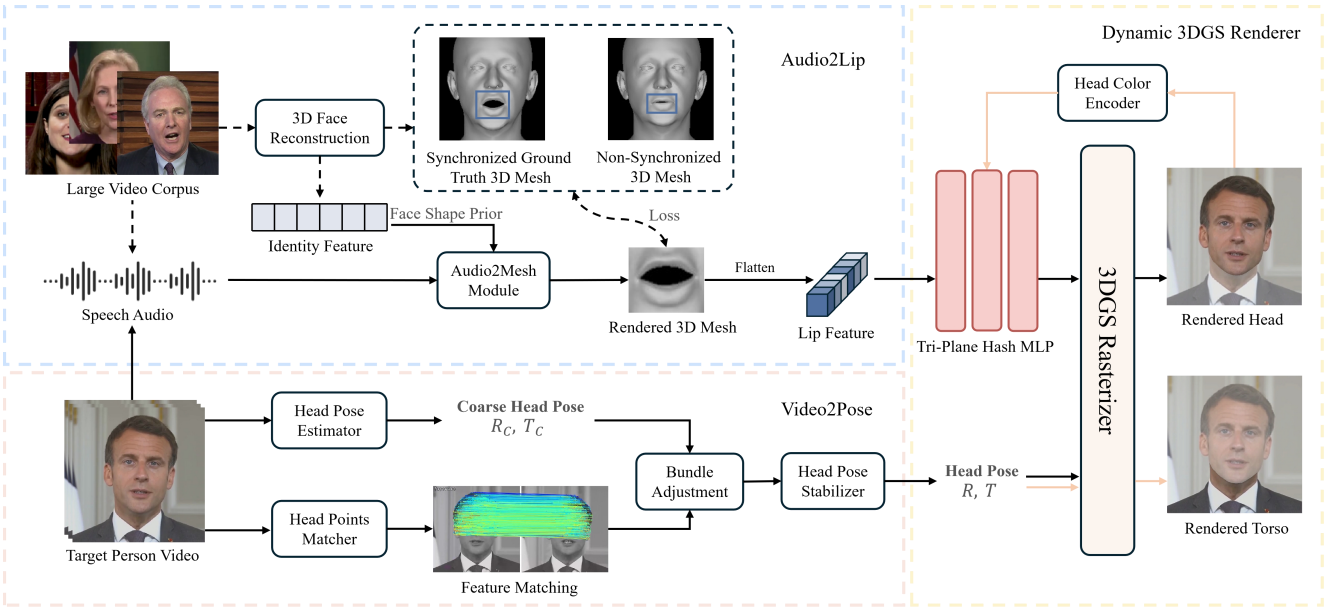

Fig. 2. Framework of the proposed VividTalker. Given a cropped reference video of a talking head and the corresponding speech, VividTalker extracts lip features and head poses using the Audio2Lip and Video2Pose. Subsequently, using tri-plane hash representation and 3DGS rasterizer, the portrait is modeled to output a speech-driven video. In the Audio2Lip, the black dashed arrow indicates that this process is only executed during the training of the Audio2Mesh Module.

3.1. Audio2Lip

3D Audiovisual Dataset. In this work, we have constructed a dataset named 3D-HDTF, derived from a large high-resolution audiovisual dataset, HDTF [44], to pretrain the Audio2Mesh module discussed subsequently. Specifically, we utilized the existing 3D face reconstruction method, SPECTRE [7], to predict the shape, expression, and pose parameters of FLAME [20]. FLAME is a parametric head model that uses linear transformations to describe identity and expression-dependent shape variations, as well as standard linear blend skinning (LBS) to model neck, jaw, and eyeball rotations. We obtained a mesh composed of thousands of 3D vertices through FLAME's forward pass, serving as 3D supervisory information. Focusing solely on the lip motion changes, we set the pose parameters for neck and eyeball rotations to zero across all video frames and discarded the 3D vertex data of non-lip regions. Additionally, we filtered out unusable samples (*e.g.*, failed face detections, misaligned audiovisual data) to improve the dataset's quality. Finally, we collected paired audio mesh sequences to create our dataset, dividing each audio mesh sequence into training and test sets at a ratio, with all sequences processed at 25 fps.

Audio2Mesh Module. Due to task limitations, existing NeRF-based and 3DGS-based methods are mostly constrained to learning the correspondence between audio and lip movements from 5-minute speech videos, making precise lip-syncing very challenging. Moreover, these methods usually utilize audio encoders specifically designed for speech recognition tasks, such as DeepSpeech [1] and HuBERT [11], to extract audio features. However, these audio encoders focus on learning the feature distribution from audio to text rather than from audio to lip movements, which results in their inability to accurately describe the actual lip motion. Finally, the intricate relationship between speech and lip makes it arduous to learn the mapping from speech to lip directly.

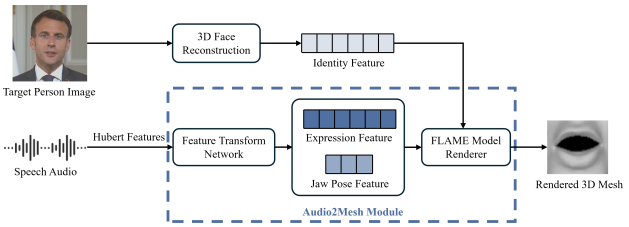

Fig. 3. The detailed structure of Audio2Mesh Module.

To tackle these issues mentioned above, we pre-trained an Audio2Mesh module on the previously constructed 3D audiovisual dataset. This module learns the mapping between the feature distribution of audio and the corresponding feature distribution of lip shapes, significantly aiding in guiding lip synthesis. The detailed structure of the Audio2Mesh module is depicted in Fig. 3. As shown in the figure, our Audio2Mesh module comprises a feature trans-

form network and a FLAME facial model renderer. For the feature transform network, we utilize a stacked one-dimensional convolutional neural network to reduce the dimensionality of HuBERT features H extracted from different audio inputs to low-dimensional features $Conv1D(H)$. These features are then decoded using multiple fully connected layers to obtain a feature vector $F \in \mathbb{R}^{53}$. Considering that the shape of the lips during movement contains the speaker's identity features, we introduce the FLAME facial model renderer to embed the identity information as a prior into the entire training process. Specifically, the feature vector F is subsequently split into two components, corresponding to the expression feature $P_{exp} \in \mathbb{R}^{50}$ and the jaw pose feature $P_{jaw} \in \mathbb{R}^3$ within the FLAME model. By incorporating these features along with the lip identity feature $P_{identity}$ extracted by the 3D face reconstruction method into the FLAME face model renderer, we can reconstruct a facial mesh sequence consisting of thousands of 3D vertices that is related to the speaker's identity. Given the differentiable nature of the entire reconstruction process, we can optimize our feature transform network through backpropagation. The reconstruction process's loss function can be formulated as follows:

$$F = Dec(Conv1D(H)) \tag{1}$$

$$F = \begin{bmatrix} P_{exp} \\ P_{jaw} \end{bmatrix} \tag{2}$$

$$\begin{aligned} \widehat{V} &= FLAME(P_{exp}, P_{jaw}, P_{identity}) \\ &= \bar{V} + B_e P_{exp} + B_j P_{jaw} + B_s P_{identity} \end{aligned} \tag{3}$$

$$L_{rec} = \frac{1}{|L|} \sum_{i \in L} \|v_i - \hat{v}_i\|_2^2 \tag{4}$$

Here, $\widehat{V} = \{\hat{v}_i\}_{i=1}^N$ and $V = \{v_i\}_{i=1}^N$ denote the predicted and actual head mesh sequences, respectively, both consisting of N three-dimensional vertices v. $\bar{V}$ represents the average head mesh sequence, while B_e, B_j and B_s represent the basis matrices of expression, jaw pose and shape respectively. Notably, L represents the set of lip vertices, meaning the reconstruction loss is defined as the mean squared error (MSE) loss computed between the lip vertices within the head mesh sequences. The mentioned V and $P_{identity}$ can be obtained via SPECTRE. To enhance the Audio2Lip module's ability to distinguish between different audio features, we introduce a contrastive loss [5] that aligns the distribution of the predicted mesh sequence $\widehat{V} = \{\hat{v}_i\}_{i=1}^N$ with the actual mesh sequence $V = \{v_i\}_{i=1}^N$. The loss function is defined as follows:

$$D_E = \sqrt{\sum_{i \in L} (v_i - \hat{v}_i)} \tag{5}$$

$$L_{cons} = Y D_E^2 + (1 - Y) \max(m - D_E, 0)^2 \tag{6}$$

Here, D_E denotes the Euclidean distance between mesh sequences, and m is a manually set threshold parameter to measure the distance between unsynchronized mesh vertices. We use V to represent the ground truth of the predicted $\widehat{V}$. In our task, Y indicates whether the audio is synchronized with the 3D facial model (with $Y = 1$ for synchronization and $Y = 0$ otherwise).

Finally, our synchronization loss function can be defined as follows:

$$L_{sync} = \omega_1 L_{rec} + \omega_2 L_{cons} \tag{7}$$

The Audio2Mesh module is trained by minimizing the above loss. Ultimately, we flattened the 3D mesh $\widehat{V}$ into a one-dimensional feature vector to be used as the lip feature extracted from the audio.

Person-Specific Fine-Tuning. In practice, even if two individuals utter the same content, their lip movements can vary due to distinct speaking styles. The aforementioned Audio2Mesh module takes into account person's identity information, but it neglects the speaking style of the target object. This makes it difficult to accurately model the lip movements. To mitigate this issue, we perform full person-specific fine-tuning on the pre-trained Audio2Mesh module to adapt it to new speaking styles. Specifically, we extract a video segment of the target subject and utilize the SPECTRE method to obtain the corresponding mesh sequence and shape features. Moreover, we introduce a motion loss L_{motion} [39] based on the synchronization loss function to ensure temporal consistency between adjacent frames, thereby eliminating jitter in the predicted mesh sequence $\widehat{V}$:

$$L_{motion} = \sum_{t=0}^{T-1} \|V_t - \widehat{V}_t\|_2^2 + \omega \sum_{t=1}^{T-1} \|V_t - V_{t-1} - (\widehat{V}_t - \widehat{V}_{t-1})\|_2^2 \tag{8}$$

Here, T denotes the manually set number of adjacent frames, and ω represents the weight used to balance the two terms. Subsequently, we further train the Audio2Mesh module based on the network weights previously learned from a large-scale corpus.

3.2. Video2Pose

Head Pose Estimator. Head pose estimation involves calculating a person's facial orientation in 3D space based on a 2D facial image, typically represented by rotation angles R and translation vectors T. Prior to estimating the head pose parameters, an optimal focal length is determined through multiple iterative calculations. The objective of each iteration is to minimize the error between the projected 3D coordinates of the 3D Morphable Model (3DMM) onto the 2D landmarks and the actual landmarks in the video frames.

To further reduce computational complexity, we generally select landmarks from a subset of video frames as supervisory information and set the focal length within a predefined range. Appropriate step sizes are then employed in the iterations to determine the most accurate focal length. The focal length f can be determined using the following formula:

$$f = \arg \min_{f_i \in [f_{\min}, f_{\max}]} L_i(P_{2D}, P_{3D}(f_i, R_i, T_i)) \quad (9)$$

Here, L_i denotes the mean squared error between the landmarks, and $P_{3D}(f_i, R_i, T_i)$ represents the landmarks projected through the 3DMM model given the focal length f_i, rotation parameters R_i, and translation parameters T_i. L_{2D} corresponds to the actual landmarks in the video frames. Once the focal length f is determined, we optimize the head pose parameters using the landmarks from all video frames. The optimization process can be formulated as follows:

$$R_{opt}, T_{opt} = \arg \min_{R,T} L(P_{2D}, P_{3D}(f_{opt}, R, T)) \quad (10)$$

The optimized head pose parameters are markedly superior to the results estimated using only a subset of video frames.

Head Feature Point Matcher. Previous approaches typically compute projection loss using a limited number of key points identified by facial keypoint detection algorithms, leading to potential pose estimation errors due to inaccuracies in keypoint detection. To enhance the accuracy of rotation parameters R and translation parameters T, we employ a feature point detection algorithm [6] to capture dense feature points, followed by image feature matching techniques [29] to track these points across different video frames. Given the possibility of incorrect matches during the matching process, we apply Random Sample Consensus (RANSAC) [8] to filter these matches. Through this module, our method aligns a greater number of more precise key points across all frames, thus improving the accuracy of head pose parameter estimation.

Bundle Adjustment. Given a set of feature points and head pose parameters, we introduce a two-stage optimization framework [9] to enhance the accuracy of head pose estimation. Initially, we randomly initialize the 3D coordinates of a series of feature points and project them onto a 2D plane to align them with the 2D feature point coordinates obtained from the Head Feature Point Matcher. The alignment between the projected feature points P and the corresponding matched feature points M is evaluated by the following function:

$$L_{coarse} = \sum_i \|P_i - M_i\|_2^2 \quad (11)$$

In the second stage, we perform a joint optimization of the 3D feature points and head pose to achieve more precise results. Utilizing gradient descent, the algorithm can automatically adjust the spatial coordinates, rotation angles R, and translation vectors T to minimize the alignment loss L_{fine}. This process is formulated as follows:

$$L_{fine} = \sum_i \|P_i(R, T) - M_i\|_2^2 \quad (12)$$

Through the two-stage optimization process, the resulting head pose parameters are significantly closer to the ground truth.

Head Pose Stabilizer. While we have obtained relatively accurate pose parameters through the Head Feature Point Matcher and Bundle Adjustment, the varying accuracy of pose parameter estimation across video frames necessitates the stabilization of head pose parameters. To address this, we employ the Savitzky-Golay filter [30] to smooth the pose parameters. This filter fits the head pose using polynomial least squares over several consecutive video frames. Through this module, our method achieves pose parameters that conform to general head motion patterns, thereby reducing head jitter and significantly enhancing the realism of talking portraits.

Head-Aware Torso Generator. To simulate torso movement, we trained a dynamic Gaussian radiance field to independently render the torso. As depicted in Fig. 2, we use the head poses obtained from the Head Pose Stabilizer as conditions to infer torso movements, aiding the model in implicitly learning torso posture variations, thereby achieving natural renderings.

Moreover, inspired by [37], we use the rendered results of the head Gaussian radiance field as input conditions for training the torso Gaussian radiance field. This approach helps alleviate the head-torso separation issue that may arise during substantial head movements. Specifically, we use the synthesized head image color C_{head} as pixel conditions for the torso radiance field. The entire rendering process for the torso can be formulated as follows:

$$\delta = THM(\mu, C_{head}) \quad (13)$$

$$(R, T, \delta) \rightarrow C_{torso} \quad (14)$$

Here, δ denotes the deformation of each Gaussian primitive, THM refers to the tri-plane hash multilayer perceptron [17], μ represents the center point of each Gaussian primitive, and $\rightarrow$ signifies the rendering process of the 3DGS rasterizer.

Table 1. Quantitative comparison under the self-reconstruction setting. The top, second-best, and third-best results are shown in red, orange, and yellow, respectively. Since Wav2Lip and DINet have access to the ground truth with the exception of the mouth region, the PSNR, LPIPS and SSIM metrics are deemed inapplicable. We achieve the best performance on most metrics.

Methods	PSNR↑	LPIPS↓	SSIM↑	FID↓	NIQE↓	BRISQUE↓	LMD↓	AUE↓	Sync-C↑	FPS↑
Ground Truth	N/A	0	1	0	0	0	0	0	8.368	N/A
Wav2Lip [28]	-	-	-	20.114	13.060	47.732	4.198	1.427	9.502	18
DINet [43]	-	-	-	10.702	12.884	45.177	3.918	1.830	6.890	21
AD-NeRF [9]	31.193	0.088	0.923	21.960	12.458	50.807	2.898	2.563	5.039	0.1
RAD-NeRF [33]	31.881	0.065	0.936	11.643	12.207	45.739	3.021	2.129	5.183	29
ER-NeRF [19]	31.550	0.036	0.931	6.785	11.943	37.549	2.845	1.889	5.588	31
GaussianTalker [4]	32.417	0.051	0.942	8.607	12.396	43.404	2.900	2.479	5.242	104
TalkingGaussian [17]	32.550	0.031	0.946	6.722	12.257	40.638	2.736	1.006	5.984	139
InsTaG [18]	30.137	0.047	0.916	10.369	12.115	40.935	2.955	1.988	6.344	116
Ours	34.332	0.025	0.962	5.724	12.284	38.746	2.535	0.838	6.524	135

4. Experiments

4.1. Experimental Settings

Dataset. To compare with previous methods, we employed the dataset from publicly-released video sets [9, 31, 37]. Specifically, this dataset consists of four high-definition speech video segments. The average length of these videos is approximately 6640 frames, with a frame rate of 25 FPS. Moreover, each video segment has been cropped to ensure the speaker remains centered in the frame, and all video resolutions have been standardized to 512×512.

Comparison Baselines. In our experiments, we compare VividTalker with the following methods: **1)** 2D generation methods based on GAN networks, including those that do not require person-specific training (Wav2Lip [28] and DINet [43]). **2)** NeRF-based methods, such as AD-NeRF [9], RAD-NeRF [33], and ER-NeRF [19], which render talking heads by training a person-specific radiance field. **3)** The most relevant 3DGS-based methods, such as GaussianTalker [4], TalkingGaussian [17], and InsTaG [18], which nicely incorporate 3DGS into talking portrait synthesis.

Implementation Details. We implemented VividTalker using the PyTorch framework and trained it on an NVIDIA RTX 4070 GPU. The pre-training and fine-tuning of Audio2Mesh take approximately 4 hours and 10 minutes respectively to converge. For the dynamic 3DGS renderer [17], we trained the face and lips of each portrait separately for 50,000 iterations, followed by a joint training for an additional 10,000 iterations. The Adam [16] and AdamW [21] optimizers were employed during training.

Table 2. Quantitative comparison under the lip-synchronization setting. We utilize two different audio samples to driven the same subject.

	Audio A		Audio B	
Methods	Sync-C↑	Sync-E↓	Sync-C↑	Sync-E↓
Ground Truth	5.999	0	5.131	0
Wav2Lip [28]	8.371	6.013	9.248	6.012
DINet [43]	5.694	8.777	6.875	8.096
AD-NeRF [9]	3.843	10.267	4.283	10.308
RAD-NeRF [33]	3.793	10.234	4.611	10.014
ER-NeRF [19]	3.887	10.167	4.547	9.980
GaussianTalker [4]	3.679	10.488	4.327	10.248
TalkingGaussian [17]	4.218	10.021	5.127	9.506
InsTaG [18]	4.442	9.249	5.526	9.004
Ours	4.471	9.883	5.737	8.960

4.2. Quantitative Evaluation

Comparison Settings and Metrics. To comprehensively evaluate head reconstruction quality, our quantitative comparison is divided into two settings: self-reconstruction and lip-synchronization. In the self-reconstruction setting, we split the four aforementioned high-definition video segments into training and test sets, using the audio, eye closure conditions, and pose sequences from the test set to reconstruct the talking portraits for quality assessment. Various image quality evaluation metrics are employed, including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [36], Learned Perceptual Image Patch Similarity (LPIPS) [40], and Frechet Inception Distance (FID) [10]. While high PSNR images may not always align with human visual perception in terms of texture details [42], we also employ two no-reference evaluation

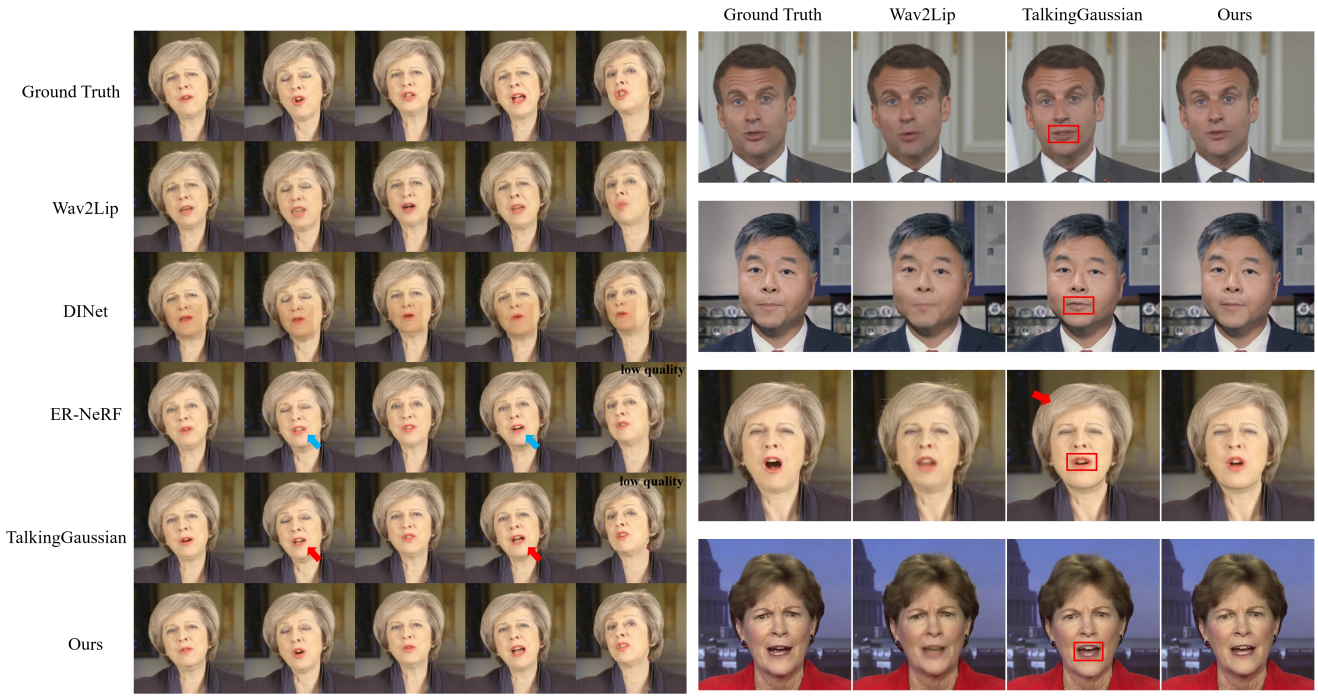

Fig. 4. Qualitative comparison of facial synthesis by different methods. Our method has the best visual effect on lip movements without the problem of separation of head and torso. Please zoom in for better visualization.

methods: Natural Image Quality Evaluator (NIQE) [24] and Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [23] for a more comprehensive measurement of image quality. Additionally, we use Landmark Distance (LMD) [3], SyncNet Confidence Score (Sync-C) [28], and Action Unit Error (AUE) [2] to evaluate the motion quality of the reconstructed portraits. LMD and Sync-C assess lip-sync accuracy, while AUE measures facial motion accuracy.

For the lip-synchronization setting, we use audio tracks from unseen videos to drive the model trained in the first setting and evaluate its lip-sync performance. This setting aims to focus on the model's performance under cross-domain audio input conditions. Specifically, we use audio samples from SynObama [32] and NVP [34] as two test audios, A and B. Due to the absence of ground truth images, we quantitatively measure lip-sync quality using non-comparative metrics, including SyncNet Confidence Score (Sync-C) and SyncNet Error Distance (Sync-E).

Evaluation Results. We present the experimental results for the two settings in Tables 1 and 2, respectively. **1)** In the self-reconstruction setting, our method achieves the best image quality and surpasses most methods in motion quality. Regarding image quality, our method excels in nearly all aspects, thanks to significant optimizations in lip movement and pose. Notably, in terms of the PSNR metric, our method improves by approximately 2 units over the previous state-of-the-art method, TalkingGaussian.

In terms of motion quality, our method surpasses existing NeRF-based methods and 3DGS-based methods across all metrics. Although one-shot generation methods like Wav2Lip and DINet excel in the Sync-C metric, they perform poorly in LMD and AUE due to the absence of person-specific training. **2)** In the lip-synchronization setting, our method maintains excellent generalization performance with different input audio tracks. With the support of the Audio2Lip, our model effectively learns lip movement patterns from high-quality and diverse audio features, thus overcoming the limitations of small-sample 3DGS methods.

Additionally, we compared frames per second (FPS) to evaluate the efficiency of each method. As shown in Table 1, our method maintains excellent performance in inference FPS, closely matching TalkingGaussian, thereby demonstrating the high efficiency of VividTalker.

4.3. Qualitative Evaluation

Evaluation Results. To more intuitively assess image quality, we present a comparison between our method and others in Fig. 4. In this figure, VividTalker demonstrates more precise facial details. Compared to Wav2Lip and DINet, our method better preserves the subject's identity while offering higher fidelity and resolution. Compared to ER-NeRF, our method avoids head-torso separation via the Head Pose Stabilizer and produces more accurate lip

Fig. 5. Qualitative comparison of facial synthesis under the lip-synchronization setting. The sequence depicts the lip shape conforming to specific phonemes in the spoken words 'said', 'on', 'homegrown', 'Orlando', 'outside', 'impression', 'journalists', and 'gradually'. Please zoom in for better visualization.

shapes. Compared to TalkingGaussian, our method excels in lip-sync performance, mainly due to the audiovisual consistency provided by the Audio2Lip. Notably, our method performs well even with large head rotations, highlighting its robustness. Overall, our method achieves the best overall visual effect.

Fig. 5 illustrates the results of the cross-driven experiments. Under two out-of-domain audio driving conditions, we select four representative keyframes to evaluate lip-sync accuracy. The 2D-based approaches (Wav2Lip and DINet) suffer from unstable synchronization, where lip movements occasionally fail to align with the corresponding phonemes. Moreover, the generated lips are either insufficiently sharp or unable to maintain subject consistency. ER-NeRF struggles to capture expressive lip dynamics for complex words, for instance failing to produce the correct mouth shape for the word 'homegrown'. TalkingGaussian and InsTaG often produce overly smoothed lip motions, with the latter occasionally exhibiting distorted lip shapes under extreme conditions. In addition, we highlight regions with severe rendering blur using red circles. In contrast to these baselines, our method achieves superior performance by jointly improving lip clarity and phoneme-to-mouth correspondence.

User Study. To better evaluate visual quality in real-world scenarios, we conducted a user study in which 28 talking head videos were generated using 7 methods. We then invited 16 participants to rate these methods based on three

Table 3. User Study. Rating is on a scale of 1-5; the higher the better. The top, second-best, and third-best results are shown in red, orange, and yellow, respectively.

Methods	Wav2Lip [28]	DINet [43]	ER-NeRF [19]	GaussianTalker [4]	TalkingGaussian [17]	InsTaG [18]	**VividTalker**
Image Quality	1.67	2.64	3.27	3.39	3.58	2.98	4.16
Lip-sync Accuracy	2.75	3.44	3.13	2.98	3.53	3.61	4.09
Video Realness	1.70	1.88	3.03	3.06	3.38	2.97	3.97

criteria: (1) Image Quality, (2) Lip-sync Accuracy, and (3) Video Realness. The results, reported in Table 3, show that our VividTalker achieves the highest scores across all three aspects, demonstrating the potential value of our method for real-world applications.

4.4. Ablation Study

In this section, we perform ablation studies to prove the necessity of each component in VividTalker. The results are shown in Table 4 and Table 5.

Audio2Lip. The Audio2Lip provides lip information synchronized with the audio to drive Gaussian primitives deformation. When this module is replaced, all metrics deteriorate, especially with a noticeable increase in LMD error. From Fig. 6 (a), we can observe a decline in the synchronization of lip movements, highlighting that generic speech recognition features fail to capture the nuanced mapping to lip kinematics as effectively as our dedicated Audio2Lip module.

Removing the FLAME renderer from Audio2Mesh leads to a performance decline, confirming that using the raw network output directly is suboptimal. The FLAME renderer is crucial as it enforces a realistic facial geometry prior and explicitly incorporates the speaker's identity feature. This ensures the predicted lip motions are both anatomically plausible and consistent with the target portrait's identity.

Further dissecting the Audio2Lip module, we evaluate two critical stages: pre-training and fine-tuning. Without pre-training, the model is deprived of the prior knowledge of generalized audio-lip correlations learned from our large-scale 3D audiovisual dataset. This results in the most significant performance drop among the Audio2Lip variants, with the LMD error increasing by approximate 10%. This underscores that pre-training provides a fundamental and robust initialization that cannot be compensated for by only learning from the limited target subject's data. Comparatively, removing the person-specific fine-tuning also degrades performance, but to a lesser extent than removing pre-training. This indicates that while adapting to the target's speaking style is important, the foundational knowledge acquired during pre-training is even more critical for achieving accurate lip synchronization.

Table 4. Ablation study on Audio2Lip. We show the PSNR, LPIPS, and LMD in different cases.

Setting	PSNR↑	LPIPS↓	LMD↓
Ours	34.332	0.025	2.535
replace Audio2Lip Module with DeepSpeech	34.007	0.026	2.705
w/o FLAME Model Renderer in Audio2Mesh	34.149	0.026	2.677
w/o pretraining	34.037	0.026	2.755
w/o fine-tuning	34.100	0.027	2.694

(a) Wrong lip sync

(b) Separation of head and torso

Fig. 6. Ablation Study on Audio2Lip and Video2Pose. Removing them will lead to (a) and (b).

Video2Pose. When the Video2Pose is removed, the image quality significantly decreases, particularly in PSNR. Fig. 6 (b) also illustrates the separation between the head and torso.

Removing the head pose stabilizer results in performance decreases, though less significantly than removing Video2Pose, indicating that preceding components (Head Feature Point Matcher and Bundle Adjustment) contribute effectively. Since the optimization of head pose parameters

Table 5. Ablation study on Video2Pose.

Setting	PSNR↑	LPIPS↓	LMD↓
Ours(Head)	34.332	0.025	2.535
w/o Video2Pose	33.151	0.029	2.594
w/o Head Pose Stabilizer	33.857	0.027	2.591
Ours(Head and Torso)	28.598	0.059	-
w/o Head-Aware	28.196	0.062	-

varies across frames, the stabilizer is essential for enforcing temporal smoothness. Without it, the generated pose sequence exhibits noticeable high-frequency jitter.

Finally, we demonstrated the necessity of head-aware features for torso reconstruction. After eliminating the head image condition, the torso-3DGS is unable to perceive the head position, resulting in head-torso separation issues.

5. Conclusion and Future Works

In this paper, we propose VividTalker, a novel talking portrait synthesis framework, which employs a highly synchronized method based on 3DGS. The framework is primarily composed of a Audio2Lip module and a Video2Pose module, which maintain the subject's identity while generating synchronized lip movements and stable body poses. Extensive experiments demonstrate that our framework achieves significant improvement comparing other advanced methods in realistic talking portrait synthesis. For future work, we plan to construct our 3D audio-visual dataset based on a larger-scale lip-reading corpus to further enhance the generalization performance of the Audio2Lip.

6. Ethical Statement

We hope our method can enhance interactive experiences and benefit humanity, but we are also cognizant of the potential risks of misuse. As part of our responsibility, we will share the generated results to promote the development of robust deepfake detectors. We believe that the responsible use of this technology will foster healthy development in machine learning research and the digital industry. We are considering protective measures, such as adding digital watermarks to real videos to prevent misuse. Furthermore, we plan to impose restrictions in the project's open-source license to prevent misuse related to "deepfakes." We hope the public will be aware of the potential risks arising from the misuse of new technologies.

Acknowledgement

This work is supported by NSFC under Grant U25A20402, the National Key R&D Program of China (No.2023YFF0615800), and the Science and Technology Funding of Sichuan Province under Grant 2024ZHCG0191 and 2026YFHZ0220.

References

[1] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR, 2016.

[2] T. Baltrušaitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, volume 6, pages 1–6. IEEE, 2015.

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

[4] K. Cho, J. Lee, H. Yoon, Y. Hong, J. Ko, S. Ahn, and S. Kim. Gaussiantalker: Real-time talking head synthesis with 3d gaussian splatting. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10985–10994, 2024.

[5] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[6] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018.

[7] P. P. Filntisis, G. Retsinas, F. Paraperas-Papantoniou, A. Katsamanis, A. Roussos, and P. Maragos. Visual speech-aware perceptual 3d facial expression reconstruction from videos. *arXiv preprint arXiv:2207.11094*, 2022.

[8] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[9] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5784–5794, 2021.

[10] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[11] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

[12] X. Ji, H. Zhou, K. Wang, Q. Wu, W. Wu, F. Xu, and X. Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

[13] X. Ji, H. Zhou, K. Wang, W. Wu, C. C. Loy, X. Cao, and F. Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14080–14089, 2021.

[14] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[15] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep video portraits. *ACM transactions on graphics (TOG)*, 37(4):1–14, 2018.

[16] D. P. Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] J. Li, J. Zhang, X. Bai, J. Zheng, X. Ning, J. Zhou, and L. Gu. Talkinggaussian: Structure-persistent 3d talking head synthesis via gaussian splatting. In *European Conference on Computer Vision*, pages 127–145. Springer, 2025.

[18] J. Li, J. Zhang, X. Bai, J. Zheng, J. Zhou, and L. Gu. Instag: Learning personalized 3d talking head from few-second video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10690–10700, 2025.

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

[20] 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.*, 36(6):194–1, 2017.

[21] I. Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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

[23] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

[24] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[25] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022.

[26] Z. Peng, W. Hu, Y. Shi, X. Zhu, X. Zhang, H. Zhao, J. He, H. Liu, and Z. Fan. Synctalk: The devil is in the synchronization for talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 666–676, 2024.

[27] Z. Peng, Y. Luo, Y. Shi, H. Xu, X. Zhu, H. Liu, J. He, and Z. Fan. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5292–5301, 2023.

[28] 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 *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492, 2020.

[29] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020.

[30] A. Savitzky and M. J. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[31] S. Shen, W. Li, Z. Zhu, Y. Duan, J. Zhou, and J. Lu. Learning dynamic facial radiance fields for few-shot talking head synthesis. In *European conference on computer vision*, pages 666–682. Springer, 2022.

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

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

[34] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 716–731. Springer, 2020.

[35] J. Wang, X. Qian, M. Zhang, R. T. Tan, and H. Li. Seeing what you said: Talking face generation guided by a lip reading expert. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14653–14662, 2023.

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

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

[38] Z. Ye, T. Zhong, Y. Ren, J. Yang, W. Li, J. Huang, Z. Jiang, J. He, R. Huang, J. Liu, et al. Real3d-portrait: One-shot realistic 3d talking portrait synthesis. *arXiv preprint arXiv:2401.08503*, 2024.

[39] C. Zhang, Y. Zhao, Y. Huang, M. Zeng, S. Ni, M. Budagavi, and X. Guo. Facial: Synthesizing dynamic talking face with implicit attribute learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3867–3876, 2021.

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

[41] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661, 2023.

[42] W. Zhang, Y. Liu, C. Dong, and Y. Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3096–3105, 2019.

[43] Z. Zhang, Z. Hu, W. Deng, C. Fan, T. Lv, and Y. Ding. Dinet: Deformation inpainting network for realistic face visually dubbing on high resolution video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3543–3551, 2023.

[44] Z. Zhang, L. Li, Y. Ding, and C. Fan. 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*, pages 3661–3670, 2021.

[45] W. Zhong, C. Fang, Y. Cai, P. Wei, G. Zhao, L. Lin, and G. Li. Identity-preserving talking face generation with landmark and appearance priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2023.