SyncGaussian: Stable 3D Gaussian-Based Talking Head Generation with Enhanced Lip Sync via Discriminative Speech Features

Ke Liu[1] , **Jiwei Wei**[1,2*] , **Shiyuan He**[1] , **Zeyu Ma**[1] , **Chaoning Zhang**[1] , **Ning Xie**[1] and **Yang Yang**[1,2]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

[2]Institute of Electronic and Information Engineering of UESTC in Guangdong, 523808

{liuke3068, mathematic6, shiyuanhe.david, chaoningzhang1990, seanxiening}@gmail.com, cnzeyuma@163.com, yang.yang@uestc.edu.cn

Abstract

Generating high-fidelity talking heads that maintain stable head poses and achieve robust lip sync remains a significant challenge. Although methods based on 3D Gaussian Splatting (3DGS) offer a promising solution via point-based deformation, they suffer from inconsistent head dynamics and mismatched mouth movements due to unstable Gaussian initialization and incomplete speech features. To overcome these limitations, we introduce SyncGaussian, a 3DGS-based framework that ensures stable head poses, enhanced lip sync, and realistic appearances with real-time rendering. SyncGaussian employs a stable head Gaussian initialization strategy to mitigate head jitter by optimizing commonly used rough head pose parameters. To enhance lip sync, we propose a sync-enhanced encoder that leverages audio-to-text and audio-to-visual speech features. Guided by a tailored cosine similarity loss function, the encoder integrates discriminative speech features through a multi-level sync adaptation mechanism, enabling the learning of an adaptive speech feature space. Extensive experiments demonstrate that SyncGaussian outperforms state-of-the-art methods in image quality, dynamic motion, and lip sync, with the potential for real-time applications.

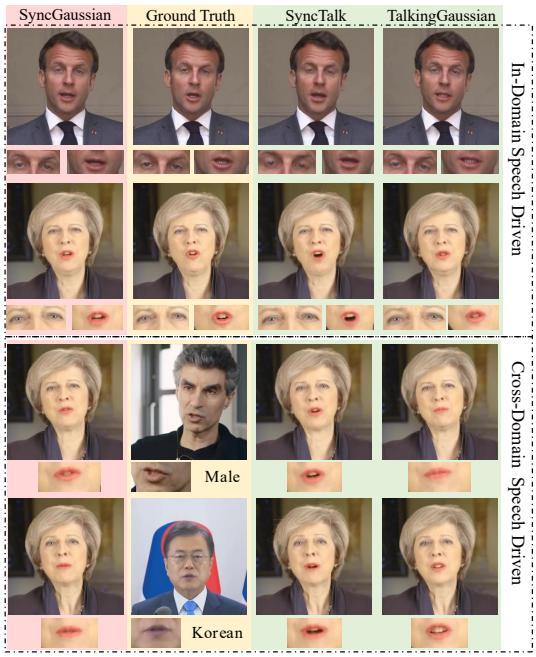

Figure 1: Our SyncGaussian outperforms NeRF and 3DGS methods in synthesizing stable and synchronized talking heads. Across both in-domain (self-reconstruction) and cross-domain (speech from different genders, languages, etc.) scenarios, our method demonstrates significant improvements in lip sync and visual quality.

1 Introduction

Given a speaker and arbitrary speech, talking head generation aims to synthesize lifelike talking heads. It has diverse applications such as digital assistants [Zhu *et al.*, 2021], animation movies [Zhong *et al.*, 2023] and video editing [Ma *et al.*, 2023]. Despite significant efforts to generate high-fidelity talking heads, ensuring stable head poses and accurate mouth movements while achieving real-time rendering speed remains a formidable challenge. This problem becomes even more pronounced in cross-domain speech driven scenarios, such as those involving different languages and genders.

Traditional generative methods based on Generative Adversarial Networks (GANs) [Wang *et al.*, 2023] excel in modeling mouth movements but struggle to maintain identity consistency across frames. Recently, emerging approaches based on Neural Radiance Fields (NeRF) [Peng *et al.*, 2024; Li *et al.*, 2023; Shen *et al.*, 2022] have succeeded in achieving photorealistic rendering by directly modifying density and color through implicit functions. Despite their ability to provide dynamic speech-lip synchronization and preserve facial details, NeRF methods suffer from slow inference speed and cannot accurately fit rapidly changing facial motions. More recently, 3D Gaussian Splatting (3DGS) [Kerbl *et al.*, 2023] has been introduced for talking head generation, significantly

* Corresponding Author (mathematic6@gmail.com)

improving rendering speed and adapting well to abrupt appearance changes. However, most 3DGS-based methods [Chen *et al.*, 2024a; Li *et al.*, 2024] consider only audio-to-text or audio-to-visual speech features. This incomplete speech condition results in mismatched mouth movements, inherently constraining the potential for achieving optimal lip sync performance. Additionally, these methods typically employ initialization strategies based on sparse 3D Morphable Models (3DMM) [Paysan *et al.*, 2009] head pose parameters. Relying on sparse parameters for head Gaussian initialization leads to unstable head poses, ultimately compromising the overall image quality.

In this paper, we introduce SyncGaussian, a real-time 3DGS-based talking head generation framework. SyncGaussian addresses the challenge of unstable head poses with a stable head Gaussian initialization strategy and proposes a sync-enhanced encoder to enhance universal lip sync performance. As shown in Figure 1, our method can generate realistic talking heads while ensuring robust lip sync across different speech-driven scenarios.

Specifically, in the stable head Gaussian initialization strategy, we use head-motion and head-points trackers to extract rough head pose parameters and dense facial keypoints. By applying a joint adjustment strategy, we achieve stable head pose parameters. Subsequently, 3DGS is applied to these parameters to initialize stable head Gaussian fields. To achieve universal lip sync across diverse speech-driven scenarios, we propose a sync-enhanced encoder that harnesses the advantages of audio-to-text speech features from Hubert [Hsu *et al.*, 2021] and audio-to-visual speech features from Audio-Visual Encoder (AVE) [Peng *et al.*, 2024]. These features are integrated into a sync adaption mechanism, facilitating the learning of an adaptive speech feature space. Then, we implement point-wise deformation utilizing Gaussian primitives and condition sets to predict the positional displacements and shapes of the talking head, thereby accurately capturing facial dynamics within complex motion fields. Finally, the deformed primitives are processed through the 3DGS rasterizer, achieving high-fidelity rendering of the target talking heads.

The main contributions are as follows:

- We propose a sync-enhanced encoder that integrates audio-to-text and audio-to-visual speech features. By employing a sync adaptation mechanism, it learns a robust speech feature space to address the problem of incomplete speech feature learning, enhancing lip sync in various speech-driven scenarios.
- We introduce a stable head Gaussian initialization strategy in Gaussian field, optimizing sparse head pose parameters to ensure smooth and stable head movements.
- Extensive experiments demonstrate that the proposed SyncGaussian can render talking heads with stable head poses in real-time while excelling in lip sync performance. All key metrics—image quality, dynamic motions, and lip sync—surpass other SOTA methods.

2 Related Work

2.1 Audio-Driven Talking Head Generation

Talking head generation aims to map acoustic features to time-aligned facial motions [Zhou *et al.*, 2021; Zhang *et al.*, 2021; Chen *et al.*, 2020], and it can learn interactive information from multi-modal spaces [Wei *et al.*, 2023; Wei *et al.*, 2021a; Liu *et al.*, 2024; Wei *et al.*, 2020; Wei *et al.*, 2021b]. Extensive research has been conducted on 2D-based methods [Du *et al.*, 2023; Prajwal *et al.*, 2020]. However, due to the absence of an explicit 3D structure, these 2D-based methods fall short in maintaining naturalness and consistency when the head pose undergoes variations.

In contrast, NeRF-based frameworks [Guo *et al.*, 2021; Liu *et al.*, 2022; Peng *et al.*, 2024] are another promising direction and have been widely explored in audio-driven talking head generation. Previous NeRF-based methods grapple with the significant computational overhead inherent in vanilla NeRF implementations. Utilizing audio as a driving force, RAD-NeRF [Tang *et al.*, 2022] and ER-NeRF [Li *et al.*, 2023] have achieved remarkable strides in both visual fidelity and operational efficiency. SyncTalk [Peng *et al.*, 2024] introduces an audio-visual encoder to bolster the generalizability of cross-domain audio inputs. However, while NeRF-based methods deliver photorealism and multi-view consistency, they struggle to smoothly represent facial motions, resulting in distorted features due to the complexity of learning discontinuous appearance changes. Compared to these methods, our SyncGaussian utilizes 3DGS to maintain accurate head structures, simplifying the learning complexity of facial motions through pure deformation representation.

2.2 3DGS-based Talking Head Generation

3D Gaussian Splatting (3DGS) [Kerbl *et al.*, 2023] presents a direct, point-based representation for radiance fields. It simplifies deformation by directly modifying a set of Gaussian primitives, enabling efficient warping of the canonical field. 3DGS introduces 3D Gaussians as a distinct discretization scheme for scene representation, enabling differentiable optimization of parameters via anisotropic splatting. Compared to traditional volume rendering of implicit neural radiance fields, 3DGS minimizes extraneous spatial computations and leverages a parallelized, visibility-centric rendering process. Some works have successfully employed 3DGS in facial animation [Chen *et al.*, 2024b], yielding promising results. In talking head generation, 3DGS has emerged as a novel technology [Li *et al.*, 2024; Chen *et al.*, 2024a; Cho *et al.*, 2024]. These methods excel in both rendering speed and synthesis quality.

However, due to the incomplete speech features and the unstable Gaussian initialization strategy, these methods suffer from head instability and limited lip sync performance. In this paper, we introduce SyncGaussian, a framework that not only maintains real-time rendering speed and synthesis quality but also effectively mitigates head jitters and significantly enhances both in-domain and cross-domain lip sync accuracy.

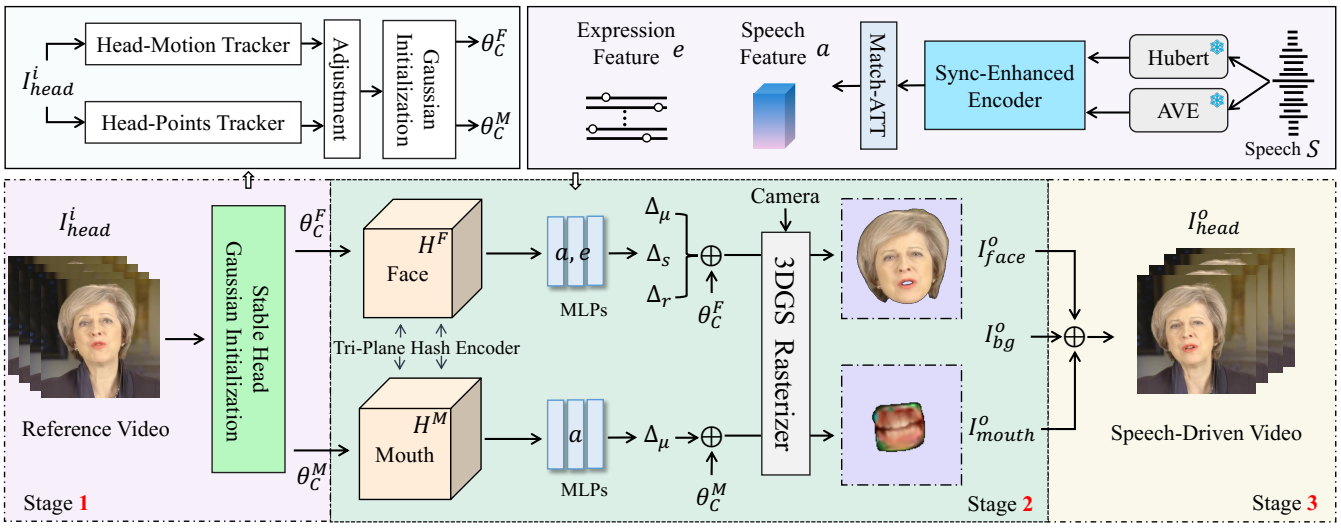

Figure 2: **Overview of SyncGaussian.** Given a reference video I^i_{head} and the speech S, SyncGaussian can extract sync-enhanced speech feature a, expression feature e, and persistent Gaussian fields parameters θ^F_C and θ^M_C. The parallel Tri-Plane Hash encoders and the MLPs jointly predict the point-wise deformation with a and e. Finally, the 3DGS rasterizer renders the modified 3D Gaussian primitives into 2D views I^o_{face} and I^o_{mouth} from the given camera. These 2D images and the background I^o_{bg} are fused to create the full talking head I^o_{head}. Stage 1 involves stable head Gaussian initialization, Stage 2 encompasses the deformation process, and Stage 3 focuses on color fine-tuning.

3 Method

3.1 Preliminaries

3D Gaussian Splatting (3DGS) [Kerbl *et al.*, 2023] employs a collection of 3D Gaussians to represent 3D information, comprising a unique set of distributions that capture spatial details effectively. Utilizing a collection of 3D Gaussian primitives θ alongside camera model parameters tailored to observing view, it computes the color C for each pixel in the rendered image. A Gaussian primitive consists of a scaling factor $s \in \mathbb{R}^3$, a rotation quaternion $q \in \mathbb{R}^4$, a mean position $\mu \in \mathbb{R}^3$, an opacity value $\alpha \in \mathbb{R}$, and a Z-dimensional color feature $f \in \mathbb{R}^Z$. Therefore, the i^{th} Gaussian primitive G_i can be represented by $\theta_i = \{\mu_i, s_i, q_i, \alpha_i, f_i\}$. With the covariance matrix $\sum$, which can be decomposed into s and q, G_i is calculated as follows:

$$G_i(x) = e^{-\frac{1}{2}(x-\mu_i)^T \sum_i^{-1}(x-\mu_i)}. \tag{1}$$

For each pixel x_p on the image, a 3DGS rasterizer is used to gather N Gaussians following the camera model to compute the color C:

$$C(x_p) = \sum_{i \in N} c_i \tilde{\alpha} \prod_{j=1}^{i-1} (1 - \tilde{\alpha_j}), \tag{2}$$

where c is the decoded color, $\tilde{\alpha}$ refers to the projected opacity. The opacity $A \in [0, 1]$ of x_p is:

$$A(x_p) = \sum_{i \in N} \tilde{\alpha_i} \prod_{j=1}^{i-1} (1 - \tilde{\alpha_j}). \tag{3}$$

3.2 Sync-Enhanced Encoder

Existing methods based on NeRF or the emerging 3D Gaussian Splatting (3DGS) consistently rely on pre-trained speech extractors such as Audio-Visual Encoder (AVE) [Peng *et al.*, 2024], DeepSpeech [Amodei *et al.*, 2016], Wav2Vec 2.0 [Baevski *et al.*, 2020], or Hubert [Hsu *et al.*, 2021].

Pre-trained on the audio-visual synchronization dataset LRS2 [Afouras *et al.*, 2018], AVE adeptly learns the feature from audio to mouth movements. While it excels at learning rich audio-to-visual features, it notably lacks in capturing audio-to-text features. This limitation significantly hinders its ability to generalize across different speech-driven scenarios. Audio-to-text speech extractors can learn general speech features, such as linguistic and prosodic features [Liu *et al.*, 2023], which are closely related to mouth movements. In Section 4.2, we demonstrate that among different audio-to-text speech extractors, Hubert performs the best in lip sync, which is consistent with the results of [Chen *et al.*, 2024a]. Therefore, as shown in Figure 3, we propose a Sync-Enhanced Encoder that adaptively incorporates discriminative speech features, enhancing the lip sync perfor-

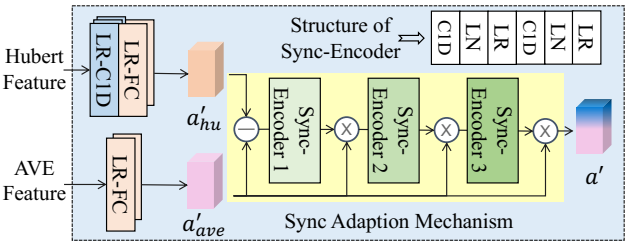

Figure 3: **The detailed structure of the Sync-Enhanced Encoder.** $\ominus$ is the subtraction operator, $\otimes$ refers to the element-wise multiplication. The sync-encoder consists of the Conv1d (C1D) layer, LayerNorm (LN) layer, and LeakyRelu (LR) activation function. FC represents the fully connected layer.

mance across different speech-driven scenarios.

Specifically, we first utilize frozen Hubert and AVE to extract speech features with different distributions from the same speech S. The learnable audio-to-text speech features, denoted as a'_{hu}, are obtained through a combination of one-dimensional convolutional layers, the LeakyReLU activation function, and fully connected layers. The learnable audio-to-visual features, denoted as a'_{ave}, are obtained using the LeakyReLU activation function and fully connected layers. Then, in the sync adaption mechanism, a subtraction operator is applied to a'_{hu} and a'_{ave} to generate discriminative features. These features are then fed into the sync-encoder, which adapts its output to a'_{ave} using an element-wise multiplication operator. To effectively integrate the discriminative feature while maintaining audio-visual consistency, this operation is performed sequentially three times. Finally, the adaptive speech feature space a is learned under the supervision of a cosine similarity loss L_{ah}:

$$L_{ah} = 1 - cos\left(a, a_{ave}\right), \tag{4}$$

where a and a_{ave} are processed by a' and a'_{ave} through the Match-ATT in Figure 2. Match-ATT [Guo *et al.*, 2021; Shen *et al.*, 2022; Tang *et al.*, 2022] is used to ensure the output dimensions align with the spatial coordinates.

3.3 Stable Head Gaussian Initialization

Previous 3DGS-based methods directly use 3DMM models to estimate head poses. However, due to the complexity of head dynamics, learning deformations directly from these sparse parameters often results in head jitter. In Figure 2, we combine the head-motion tracker and head-points tracker to generate stable head poses following [Peng *et al.*, 2024], a technique utilized in NeRF. Based on this, we use vanilla 3D Gaussian Splatting (3DGS) to initialize a stable head Gaussian field.

To initially estimate the head pose, we iteratively determine the optimal focal length within a predefined range, requiring i iterations. For each focal length candidate foc_i, we re-initialize the rotation and translation values. The optimal focal length foc_{opt} is obtained by minimizing the error between the landmarks of the reference video and the projected landmarks extracted by 3DMM, using mean squared error as the optimization criterion. Once foc_{opt} is determined, the head-motion tracker refines the rotation and translation parameters across all frames to achieve closer alignment between the projected and actual landmarks. To enhance the accuracy of head pose parameters, we use the head-points tracker based on an optical flow estimation model for tracking facial keypoints. After acquiring the facial motion optical flow, we apply the Laplacian filter to identify the keypoints with the most pronounced flow variations. Subsequently, we meticulously track the motion trajectories of these identified keypoints within the flow sequence, ensuring a precise capture of their dynamic behavior.

Given the facial keypoints and head pose, we refine their accuracy through a two-stage optimization approach adapted from [Guo *et al.*, 2021]. In the first stage, we optimize the position of randomly initialized 3D coordinates by minimizing the L_2 loss function between projected keypoints and tracked keypoints. In the second stage, we refine the 3D keypoints and jointly optimize the associated head pose parameters. The algorithm adjusts the spatial coordinates, rotation angles, and translations to minimize the alignment error using the other L_2 loss function. Finally, the generated head pose is smooth and stable, making it suitable for 3DGS to initialize the stable Gaussian fields θ_C^F and θ_C^M.

3.4 3DGS for Talking Head Generation

We utilize stable head Gaussian fields in conjunction with grid-based motion fields to achieve deformations within the Gaussian radiance field, effectively portraying diverse head motions in 3D space. As shown in Figure 2, the canonical parameters $\theta_C = \left\{\theta_C^F, \theta_C^M\right\}$ can preserve the persistent Gaussian primitive. The coarse mean field θ_C is initialized through 3DGS using stable head pose parameters. Since Gaussian primitives lack a regional position encoding for a fully explicit spatial structure, we adopt the efficient tri-plane hash encoder $H = \left\{H^F, H^M\right\}$ and MLPs for position encoding. Given the input center μ_i, the motion field predicts a point-wise deformation $\delta_i = \left\{\delta_i^F, \delta_i^M\right\}$ for each primitive, representing the motion without being influenced by color or opacity changes. δ_i can be calculated by:

$$\delta_i = MLP\left(H\left(\mu_i\right) \oplus C\right), \tag{5}$$

where $\oplus$ is concatenation and C is the condition feature. Ultimately, the deformed Gaussian primitives are generated from face and mouth motion fields, utilizing the 3DGS rasterizer to render the talking head.

To overcome the gradient vanishing issue during the learning of deformations, which arises when the target primitive position is too far from the predicted results, we adopt an incremental sampling strategy from [Li *et al.*, 2024]. Based on the detected action units and landmarks, we utilize m to quantify how much each facial motion deviates from its original or expected state. During the k^{th} training iteration, we leverage a sliding window to select a vital training frame at position j. The motion metric, m_j, adheres to a predefined condition, $m_j \in [B_l + k \times T, B_u + k \times T]$, where B_l is the initial lower bound of the sliding window, B_u is the upper bound, and T refers to the step length.

Due to the granularity problem caused by the motion inconsistency between the face and the inside mouth in the grid-based motion fields, similar to [Li *et al.*, 2024], we use two branches, i.e., face branch and mouth branch, to predict the deformation respectively. These two branches are obtained using semantic masks from the BiSeNet [Yu *et al.*, 2018] parser and EasyPortrait [Kvanchiani *et al.*, 2023]. In the face branch, we incorporate a region attention mechanism [Li *et al.*, 2023] within the grid-based motion fields to enhance the learning process for conditional deformation. This process is guided by the speech condition a and the upper-face expression condition e. In the mouth branch, we only predict the translation $\triangle\mu_i$ of the i^{th} primitive using the speech condition a. This simplified design is due to the relatively simple motion patterns of the mouth, which are primarily correlated with speech. The rendered face, mouth, and the background are fused to generate the talking head. Therefore, the talking head color C_{head} of each pixel x_p can be represented as:

Methods	Image Quality			Dynamic Motion		Lip Sync	
	PSNR ↑	LPIPS ↓	SSIM ↑	LMD ↓	AUE-(L/U) ↓	Sync-C ↑	Sync-E ↓
Ground Truth	N/A	0.000	1.000	0.000	0.000/0.000	8.192	6.307
Wav2Lip [Prajwal *et al.*, 2020]	33.60	0.0661	0.903	3.089	0.718/-	**9.429**	**5.245**
VideoReTalking [Cheng *et al.*, 2022]	31.48	0.0546	0.816	3.765	0.764/-	6.930	7.025
TalkLip [Wang *et al.*, 2023]	33.16	0.0697	<u>0.904</u>	2.929	<u>0.529</u>/-	4.482	8.217
ER-NeRF [Li *et al.*, 2023]	30.69	0.0617	0.856	3.039	0.762/1.021	5.720	8.295
SyncTalk [Peng *et al.*, 2024]	<u>34.95</u>	0.0312	0.903	<u>2.667</u>	0.567/<u>0.234</u>	7.130	7.134
TalkingGaussian [Li *et al.*, 2024]	33.52	<u>0.0267</u>	0.901	2.712	0.626/0.277	5.363	8.452
Ours	**35.21**	**0.0204**	**0.910**	**2.499**	**0.458/0.189**	<u>7.446</u>	<u>6.787</u>

Table 1: **Quantitative results of the in-domain speech driven setting.** The best and second-best methods are in **bold** and <u>underline</u>.

$$C_{head}\left(x_p\right) = C_{face}\left(x_p\right) \times A_{face}\left(x_p\right) + C_{mouth}\left(x_p\right)\left(1 - A_{face}\left(x_p\right)\right), \quad (6)$$

where C_{face} is the predicted face color, A_{face} refers to the opacity, and C_{mouth} represents the predicted mouth color.

3.5 Training Objectives

The same as the basic 3DGS optimization strategies, we train our model in three stages. The first two stages are applied to the face and mouth branches respectively, and the last stage is dedicated to the fused talking head.

In the initialization stage, we employ a pixel-wise L_1 loss function along with a D-SSIM term to quantify the error between the rendered image I_C, which is generated using the parameters θ_C, and the corresponding masked ground-truth image I_{mask} for each individual branch:

$$L_C = L_1\left(I_C, I_{mask}\right) + \lambda_1 L_{D-SSIM}\left(I_C, I_{mask}\right). \quad (7)$$

In the deformation stage, we use the deformed parameters as the input for the 3DGS rasterizer to render the output I_D:

$$L_D = L_1\left(I_D, I_{mask}\right) + \lambda_1 L_{D-SSIM}\left(I_D, I_{mask}\right) + \lambda_2 L_{ah}. \quad (8)$$

Ultimately, a color fine-tuning stage is undertaken to optimize the talking head. We calculate the reconstruction loss between the fused image I^o_{head} and the ground-truth video frame I^i_{head} with pixel-wise L_1 loss, D-SSIM, and LPIPS terms:

$$L_F = L_1\left(I^o_{head}, I^i_{head}\right) + \lambda_1 L_{D-SSIM}\left(I^o_{head}, I^i_{head}\right) + \lambda_3 L_{LPIPS}\left(I^o_{head}, I^i_{head}\right), \quad (9)$$

where $\lambda = \{\lambda_1, \lambda_2, \lambda_3\}$ is the weighting factors for different loss functions. Here, λ_1= 0.2, λ_2 = 0.5, and λ_3 = 0.5. During color optimization, we only update the color parameter f within the set of parameters θ_C, while suspending the densification strategy employed by 3DGS to ensure stability and precision in the color adjustment process.

4 Experiment

4.1 Experimental Settings

Dataset. To make a fair comparison, we collect four video sequences from [Li *et al.*, 2024; Ye *et al.*, 2023; Peng *et al.*, 2024], including English and French. These videos consist of one female portrait *May*, and three male portraits *Obama, Lieu, Macron*. The average length of these video clips is approximately 6,500 frames in 25 FPS. Three videos *May, Macron, Lieu* are resized to 512×512, while *Obama* is resized to 450×450.

Methods	Audio A		Audio B	
	Sync-C ↑	Sync-E ↓	Sync-C ↑	Sync-E ↓
VideoReTalking	7.264	7.394	7.480	7.652
ER-NeRF	4.111	9.906	4.260	10.062
SyncTalk	<u>7.480</u>	<u>7.319</u>	<u>7.486</u>	<u>7.367</u>
TalkingGaussian	5.097	8.928	5.764	9.107
Ours	**7.898**	**6.702**	**7.871**	**7.044**

Table 2: **Results of the cross-domain speech driven setting.** The best and second-best methods are in **bold** and <u>underline</u>.

Comparison Baselines. We compare our SyncGaussian with NeRF-based methods, including ER-NeRF [Li *et al.*, 2023], SyncTalk [Peng *et al.*, 2024], and GAN-based methods, such as Wav2Lip [Prajwal *et al.*, 2020], VideoReTalking [Cheng *et al.*, 2022], and TalkLip [Wang *et al.*, 2023], as well as a 3DGS-based method, TalkingGaussian [Li *et al.*, 2024]. The radiation field-based methods are person-specific and the GAN-based methods are person-generic.

Methods	Chinese	Japanese	Italian
SyncTalk	<u>4.992</u>	<u>6.897</u>	<u>6.329</u>
TalkingGaussian	3.967	5.439	5.675
Ours	**7.221**	**7.430**	**7.642**

Table 3: **Exploration of cross-language settings.** We report **Sync-C** (higher is better) to show the lip sync accuracy. The best and second-best methods are in **bold** and <u>underline</u>.

Implementation Details. We use PyTorch to train our model, and take Adam and AdamW as optimizers. For a specific portrait, we train face and mouth branches for 60,000 iterations, then jointly fine-tune 15,000 iterations. All experiments are performed on a NVIDIA RTX A6000 GPU. The

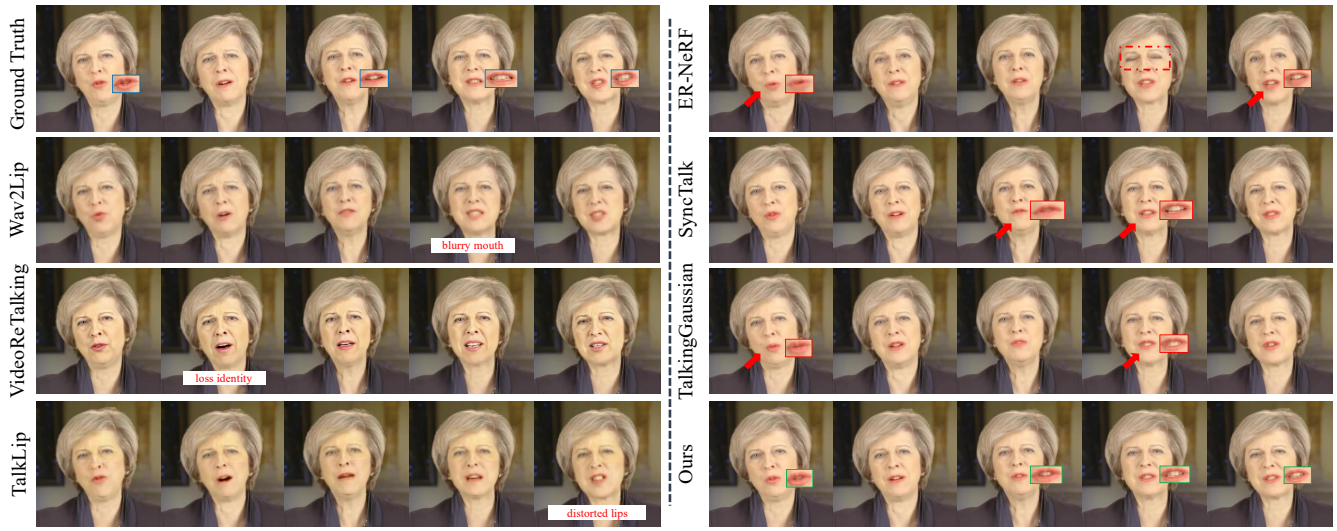

Figure 4: **Qualitative comparison under in-domain speech driven setting.** Please zoom in for better visualization.

average inference speed is 105 FPS and the overall training time is around 0.6 hours.

4.2 Quantitative Evaluation

Evaluation Metric. To comprehensively evaluate the performance of our method, we employ three distinct sets of evaluation metrics: (1) PSNR, LPIPS [Zhang *et al.*, 2018], and SSIM [Wang *et al.*, 2004] for *image quality*; (2) LMD [Chen *et al.*, 2018], upper-face action unit error (AUE-U) and lower-face action unit error (AUE-L) extracted by OpenFace [Baltrušaitis *et al.*, 2016] for *dynamic motion*; (3) The confidence score (Sync-C) and error distance (Sync-E) of SyncNet [Chung and Zisserman, 2017] for *lip sync* accuracy.

Speech Extractors	Sync-C ↑	Sync-E ↓
DeepSpeech	5.612	8.452
Wav2Vec 2.0	5.776	8.129
Hubert	**6.080**	**7.926**

Table 4: **Exploring different audio-to-text speech extractors**. The best results are in **bold**.

Comparison Settings. Our quantitative comparison contains two settings, the in-domain speech driven setting and the cross-domain speech driven setting. In the first setting, we divide each of the four videos into training ($\frac{9}{10}$) and test ($\frac{1}{10}$) sets, employing the audio, expression, and pose sequences from the unseen test set to autonomously reconstruct the talking head. In the second setting, we take the audio clips A and B from two other videos [Li *et al.*, 2024] to drive the model *May* trained in the first setting and evaluate the lip sync accuracy. Given that both Audio A and B originate from unseen videos featuring male voices, the evaluation results, notably for *May* with a contrasting gender, effectively demonstrate the generalization capability. Considering that Audio A and B match the training language of *May*, we collect Chinese, Japanese, and Italian speech to evaluate SyncGaussian in more challenging cross-domain scenarios.

Methods	Image Quality	Motion Quality	Lip Sync
VideoReTalking	2.83	3.65	2.99
ER-NeRF	3.26	3.49	1.97
SyncTalk	4.22	4.10	4.31
TalkingGaussian	4.19	3.98	3.97
Ours	**4.49**	**4.14**	**4.52**

Table 5: **User study.** The best results are in **bold**.

Evaluation Results. In Table 1, we show the results under the in-domain speech driven setting. GAN-based methods only restores the lower half of the face, we do not report AUE-U. By utilizing deformation-based motion representation and introducing a stable head Gaussian initialization strategy, we surpass all other GAN, NeRF, and 3DGS-based methods in terms of image quality and dynamic motion. As for lip sync accuracy, our SyncGaussian outperforms most methods. Wav2Lip excels in lip sync, but its inability to retain individual speaking styles results in poor dynamic motion and image quality. Compared to the latest NeRF method SyncTalk and 3DGS method TalkingGaussian, we have taken the lead in all aspects. Although our method does not lead by a large margin compared to SyncTalk, in terms of talking head rendering speed, our method is approximately 2.19 times faster (105 FPS/48 FPS).

The cross-domain speech driven setting results are shown in Table 2 and Table 3. It can be seen that our method presents the best lip sync. This demonstrates that with the assistance of sync-enhanced encoder, SyncGaussian is able to overcome the limitation of specific person data and effectively adapt to cross-domain speech-driven scenarios. SyncTalk does show some advantages, but its neglect of audio-to-text speech feature obviously lowers its effectiveness. Additionally, in Ta-

ble 4, we compare different audio-to-text speech extractors and observe that Hubert consistently performs the best. This observation underscores its selection for the sync-enhanced encoder.

Methods	PSNR ↑	LMD ↓	Sync-C ↑
w/o stable initialization	33.66	2.508	6.955
w/o a_{ave}	35.07	2.597	6.080
w/o a_{hu}	35.13	2.546	7.222
w/o sync adaption	35.06	2.547	6.974
w/o L_{ah}	35.19	2.514	7.270
SyncGaussian	**35.21**	**2.499**	**7.446**

Table 6: **Ablation study** of our contributions under the in-domain speech driven setting. The best are in **bold.**

4.3 Qualitative Evaluation

Evaluation Results. In Figure 4, we show a comparison between our method and other methods to intuitively evaluate generative quality. It can be seen that Wav2Lip, VideoReTalking and TalkLip have problems such as blurring, loss of identity, and distorted lips. This is due to the limitations of GAN and the trade-offs they made to obtain one or few-shot capabilities. Compared with ER-NeRF, our method can accurately control the eye expressions and present clearer mouth shapes. Compared to SyncTalk and TalkingGaussian, we are able to generate better mouth details and achieve superior performance in lip sync.

In Figure 5, we visualize different talking heads to facilitate a more intuitive comparison. It can be observed that compared to SyncTalk and TalkingGaussian, the advantages of our method are primarily manifested in: (1) more complete mouth details, compared to the missing teeth highlighted in the red box; (2) fuller lip shapes, as exemplified by the insufficiently opened lips in the green box; (3) clearer mouth structures, including the blurred mouth and incomplete teeth in the yellow box; and (4) more accurate mouth movements, as demonstrated by the incorrectly opened mouth in the purple box, which our method corrects.

These advantages demonstrate that our method significantly enhances the accuracy, detail preservation, and lip sync performance of speech-driven talking head generation.

User Study. In Table 5, we design a user study based on Mean Opinion Score (MOS) to provide a more comprehensive evaluation of the proposed method. We use a total of 20 videos generated by 5 methods. We invite 10 attendees to rank 5 methods, using a scale where 5 represents the best and 1 represents the worst. It can be seen that SyncGaussian performs best on all aspects, this result is consistent with both quantitative and qualitative experimental results.

4.4 Ablation Study

In Table 6, we conduct an ablation study in in-domain speech driven setting to prove the effectiveness of our contributions. We select PSNR, LMD and Sync-C for evaluation and report the average results of the four subjects.

Figure 5: **Qualitative results on different talking heads.** Please zoom in for better visualization.

When the stable head Gaussian initialization is removed, the PSNR drops significantly, with a large decline in Sync-C. This indicates that this strategy can effectively stabilize head movements, thereby improving visual quality and benefiting lip sync. Using only a_{hu} (obtained by feeding a'_{hu} into Match-ATT), i.e., without a_{ave}, to train the model results in poor performance of Sync-C and LMD, indicating that the audio-to-visual speech features is capable of extracting accurate mouth movement features. Without using a_{hu}, the overall performance is enhanced compared to without a_{ave}. Additionally, w/o L_{ah} achieves a comprehensive improvement in performance over the former, suggesting that the Hubert features and L_{ah} can bolster the model performance and generalization capabilities. In addition, we directly subtract two kinds of speech features and input them into the grid-based motion field. It can be found that although the a_{ave} is included, the performance of the model on lip sync is significantly dropped. This indicates that the sync adaption mechanism can effectively integrate audio-to-text speech features into the distribution of audio-visual consistency.

5 Conclusion

In this paper, we have introduced SyncGaussian, a 3DGS-based framework for realistic talking head synthesis with universal lip sync and stable head poses. With a stable head Gaussian initialization strategy, a sync-enhanced encoder, and a tailored cosine similarity loss, our method can overcome the head jitter problem and show robust lip sync in both cross-domain and in-domain speech driven scenarios. Through comprehensive quantitative and qualitative experiments, SyncGaussian has demonstrated superior performance in generating stable and precisely synchronized talking heads, surpassing existing methods.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62306067, and Grant 62220106008, in part by Sichuan Science and Technology Program under Grant 2024NSFSC1463, in part by Guangdong Basic and Applied Basic Research Foundation under grant No. 2025A1515010108, in part by Sichuan Province Innovative Talent Funding Project for Postdoctoral Fellows under Grant BX202405.

References

[Afouras *et al.*, 2018] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8717–8727, 2018.

[Amodei *et al.*, 2016] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang 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.

[Baevski *et al.*, 2020] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[Baltrušaitis *et al.*, 2016] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE, 2016.

[Chen *et al.*, 2018] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European conference on computer vision (ECCV)*, pages 520–535, 2018.

[Chen *et al.*, 2020] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *European conference on computer vision*, pages 35–51. Springer, 2020.

[Chen *et al.*, 2024a] Bo Chen, Shoukang Hu, Qi Chen, Chenpeng Du, Ran Yi, Yanmin Qian, and Xie Chen. Gstalker: Real-time audio-driven talking face generation via deformable gaussian splatting. *arXiv preprint arXiv:2404.19040*, 2024.

[Chen *et al.*, 2024b] Yufan Chen, Lizhen Wang, Qijing Li, Hongjiang Xiao, Shengping Zhang, Hongxun Yao, and Yebin Liu. Monogaussianavatar: Monocular gaussian point-based head avatar. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–9, 2024.

[Cheng *et al.*, 2022] Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, and Nannan Wang. Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022.

[Cho *et al.*, 2024] Kyusun Cho, Joungbin Lee, Heeji Yoon, Yeobin Hong, Jaehoon Ko, Sangjun Ahn, and Seungryong Kim. Gaussiantalker: Real-time high-fidelity talking head synthesis with audio-driven 3d gaussian splatting. *arXiv preprint arXiv:2404.16012*, 2024.

[Chung and Zisserman, 2017] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 87–103. Springer, 2017.

[Du *et al.*, 2023] Chenpeng Du, Qi Chen, Tianyu He, Xu Tan, Xie Chen, Kai Yu, Sheng Zhao, and Jiang Bian. Dae-talker: High fidelity speech-driven talking face generation with diffusion autoencoder. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4281–4289, 2023.

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

[Hsu *et al.*, 2021] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman 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.

[Kerbl *et al.*, 2023] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[Kvanchiani *et al.*, 2023] Karina Kvanchiani, Elizaveta Petrova, Karen Efremyan, Alexander Sautin, and Alexander Kapitanov. Easyportrait–face parsing and portrait segmentation dataset. *arXiv preprint arXiv:2304.13509*, 2023.

[Li *et al.*, 2023] Jiahe Li, Jiawei Zhang, Xiao Bai, Jun Zhou, and Lin 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.

[Li *et al.*, 2024] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Talkinggaussian: Structure-persistent 3d talking head synthesis via gaussian splatting. In *Proceedings of the European conference on computer vision (ECCV)*, 2024.

[Liu *et al.*, 2022] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-aware implicit neural audio-driven video portrait generation. In *European conference on computer vision*, pages 106–125. Springer, 2022.

[Liu *et al.*, 2023] Ke Liu, Dekui Wang, Dongya Wu, and Jun Feng. Speech emotion recognition via two-stream

pooling attention with discriminative channel weighting. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Liu *et al.*, 2024] Jinyuan Liu, Guanyao Wu, Zhu Liu, Long Ma, Risheng Liu, and Xin Fan. Where elegance meets precision: towards a compact, automatic, and flexible framework for multi-modality image fusion and applications. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 1110–1118, 2024.

[Ma *et al.*, 2023] Yifeng Ma, Suzhen Wang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, Zhidong Deng, and Xin Yu. Styletalk: One-shot talking head generation with controllable speaking styles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1896–1904, 2023.

[Paysan *et al.*, 2009] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009.

[Peng *et al.*, 2024] Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Jun He, Hongyan Liu, and Zhaoxin 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.

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

[Shen *et al.*, 2022] Shuai Shen, Wanhua Li, Zheng Zhu, Yueqi Duan, Jie Zhou, and Jiwen Lu. Learning dynamic facial radiance fields for few-shot talking head synthesis. In *European conference on computer vision*, pages 666–682. Springer, 2022.

[Tang *et al.*, 2022] Jiaxiang Tang, Kaisiyuan Wang, Hang Zhou, Xiaokang Chen, Dongliang He, Tianshu Hu, Jingtuo Liu, Gang Zeng, and Jingdong Wang. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. *arXiv preprint arXiv:2211.12368*, 2022.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[Wang *et al.*, 2023] Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T Tan, and Haizhou 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.

[Wei *et al.*, 2020] Jiwei Wei, Xing Xu, Yang Yang, Yanli Ji, Zheng Wang, and Heng Tao Shen. Universal weighting metric learning for cross-modal matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13005–13014, 2020.

[Wei *et al.*, 2021a] Jiwei Wei, Xing Xu, Zheng Wang, and Guoqing Wang. Meta self-paced learning for cross-modal matching. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3835–3843, 2021.

[Wei *et al.*, 2021b] Jiwei Wei, Yang Yang, Xing Xu, Xiaofeng Zhu, and Heng Tao Shen. Universal weighting metric learning for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6534–6545, 2021.

[Wei *et al.*, 2023] Jiwei Wei, Yang Yang, Xing Xu, Jingkuan Song, Guoqing Wang, and Heng Tao Shen. Less is better: Exponential loss for cross-modal matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):5271–5280, 2023.

[Ye *et al.*, 2023] 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 *The Eleventh International Conference on Learning Representations*, pages 1–15, 2023.

[Yu *et al.*, 2018] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver 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.

[Zhang *et al.*, 2021] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie 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.

[Zhong *et al.*, 2023] Weizhi Zhong, Chaowei Fang, Yinqi Cai, Pengxu Wei, Gangming Zhao, Liang Lin, and Guanbin 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.

[Zhou *et al.*, 2021] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4176–4186, 2021.

[Zhu *et al.*, 2021] Hao Zhu, Huaibo Huang, Yi Li, Aihua Zheng, and Ran He. Arbitrary talking face generation via attentional audio-visual coherence learning. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 2362–2368, 2021.