# TalkingGaussian: Structure-Persistent 3D Talking Head Synthesis via Gaussian Splatting

Jiahe Li[1], Jiawei Zhang[1], Xiao Bai[1⋆], Jin Zheng[1], Xin Ning[2], Jun Zhou[3], and Lin Gu[4,5]

[1] School of Computer Science and Engineering, State Key Laboratory of Complex & Critical Software Environment, Jiangxi Research Institute, Beihang University
[2] Institute of Semiconductors, Chinese Academy of Sciences
[3] School of Information and Communication Technology, Griffith University
[4] RIKEN AIP
[5] The University of Tokyo

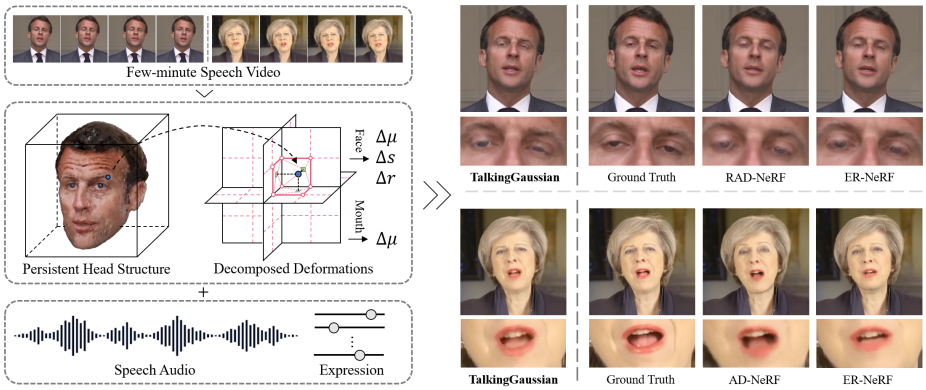


**Fig. 1:** Inaccurate predictions of the rapidly changing appearance often produce distorted facial features in previous NeRF-based methods. By keeping a persistent head structure and predicting deformation to represent facial motion, our TalkingGaussian outperforms previous methods in synthesizing more precise and clear talking heads.

**Abstract.** Radiance fields have demonstrated impressive performance in synthesizing lifelike 3D talking heads. However, due to the difficulty in fitting steep appearance changes, the prevailing paradigm that presents facial motions by directly modifying point appearance may lead to distortions in dynamic regions. To tackle this challenge, we introduce TalkingGaussian, a deformation-based radiance fields framework for high-fidelity talking head synthesis. Leveraging the point-based Gaussian Splatting, facial motions can be represented in our method by applying smooth and continuous deformations to persistent Gaussian primitives, without requiring to learn the difficult appearance change like previous methods. Due to this simplification, precise facial motions can be synthesized while keeping a highly intact facial feature. Under such a deformation paradigm, we further identify a face-mouth motion inconsistency that

⋆ Corresponding author: Xiao Bai (baixiao@buaa.edu.cn).

would affect the learning of detailed speaking motions. To address this conflict, we decompose the model into two branches separately for the face and inside mouth areas, therefore simplifying the learning tasks to help reconstruct more accurate motion and structure of the mouth region. Extensive experiments demonstrate that our method renders high-quality lip-synchronized talking head videos, with better facial fidelity and higher efficiency compared with previous methods.

**Keywords:** talking head synthesis · 3D Gaussian Splatting

## 1 Introduction

Synthesizing audio-driven talking head videos is valuable to a wide range of digital applications such as virtual reality, film-making, and human-computer interaction. Recently, radiance fields like Neural Radiance Fields (NeRF) [31] have been adopted by many methods [5, 15, 24, 36, 40, 43, 52] to improve the stability of 3D head structure while providing photo-realistic rendering, which has achieved great success in synthesizing high-fidelity talking head videos.

Most of these NeRF-based approaches [15, 24, 36, 43, 52] synthesize different face motions by directly modifying color and density with neural networks, predicting a temporary condition-dependent appearance for each spatial point in the radiance fields whenever receiving a condition feature. This appearance-modification paradigm enables previous methods to achieve dynamic lip-audio synchronization in a fixed space representation. However, since even neighbor regions can also show significantly different colors and various structures on a human face, it's challenging for these continuous and smooth neural fields to accurately fit the rapidly changing appearance to represent facial motions, which may lead to some heavy distortions on the facial features like a messy mouth and transparent eyelids, as shown in Fig. 1.

In this paper, we propose TalkingGaussian, a deformation-based talking head synthesis framework, that attempts to utilize the recent 3D Gaussian Splatting (3DGS) [20] to address the facial distortion problem in existing radiance-fields-based methods. The core idea of our method is to represent complex and fine-grained facial motions with several individual smooth deformations to simplify the learning task. To achieve this goal, we first obtain a persistent head structure that keeps an unchangeable appearance and stable geometry with 3DGS. Then, motions can be precisely represented just by the deformation applied to the head structure, therefore eliminating distortions produced from inaccurately predicted appearance, and leading to better facial fidelity while synthesizing high-quality talking heads.

Specifically, we represent the dynamic talking head with a 3DGS-based Deformable Gaussian Field, consisting of a static Persistent Gaussian Field and a neural Grid-based Motion Field to decouple the persistent head structure and dynamic facial motions. Unlike previous continuous neural-based backbones [24, 31, 32], 3DGS provides an explicit space representation by a definite set of Gaussian primitives, enabling us to obtain a more stable head structure and

accurate control of spatial points. Based on this, we apply a point-wise deformation, which changes the position and shape of each primitive while persisting its color and opacity, to represent facial motions via the motion fields. Then the deformed primitives are input into the 3DGS rasterizer to render the target images. To facilitate the smooth learning for a target facial motion, we introduce an incremental sampling strategy that utilizes face action priors to schedule the optimization process of deformation.

In the Deformable Gaussian Fields, we further decompose the entire head as a face branch and an inside mouth branch to solve the motion inconsistency between these two regions, which hugely improves the synthesis quality both in static structure and dynamic performance. Since the motions of the face and inside mouth are not related totally in tight and may be much different sometimes, it is hard to accurately represent these delicate but conflicted motions with just one single motion field. To simplify the learning of both these two distinct motions, we divide these two regions in 2D input images with a semantic mask, and build two model branches to represent them individually. As the motion in each branch has been simplified to become smooth, our method can achieve better visual-audio synchronization and reconstruct a more accurate mouth structure.

The main contributions of our paper are summarized as follows:

- We present a novel deformation-based framework that synthesizes talking heads by applying deformations to a persistent head structure, to escape an inherent facial distortion problem from the inaccurate prediction of changing appearance, enabling the generating of precise and intact facial details.
- We propose a Face-Mouth Decomposition module to facilitate motion modeling via decomposing conflicted learning tasks for deformation, therefore providing accurate mouth reconstruction and lip synchronization.
- Extensive experiments show that the proposed TalkingGaussian renders realistic lip-synchronized talking head videos with high visual quality, generalization ability, and efficiency, outperforming state-of-the-art methods on both objective evaluation and human judgment.

## 2 Related Work

**Talking Head Synthesis.** Driving talking heads by arbitrary input audio is an active research topic, aiming to reenact the specific person to generate highly audio-visual consistent videos. Early methods based on 2D generative models synthesize audio-synchronized lip motions for a given facial image [7, 12, 18, 37, 48]. Later advancements [29, 44, 46, 55] incorporate intermediate representations like facial landmarks and morphable models for better control, but suffer from errors and information loss during the intermediate estimation. Due to the lack of an explicit 3D structure, these 2D-based methods are short in keeping the naturalness and consistency when the head pose changes.

Recently, Neural Radiance Fields (NeRF) [31] has been introduced as a 3D representation of the talking head structure, providing photorealistic rendering

and personalized talking style via person-specific training. Earlier NeRF-based works [15,27,40] suffer from the expensive cost of vanilla NeRF. Successfully driving efficient neural fields [4,32] with audio, RAD-NeRF [43] and ER-NeRF [24] have gained tremendous improvements in both visual quality and efficiency. To improve the generalizability of cross-domain audio inputs, GeneFace [52] and SyncTalk [36] pre-train the audio encoder with large audio-visual datasets. However, most of these methods represent facial motions by changing the appearance of each sampling point, which burdens the network with learning the jumping and unsmooth appearance changes, resulting in distorted facial features. Although some works [25, 40, 53] have introduced a pre-trained deformable fields module for few-shot settings, the lack of fine-grained point control and precise head structure brings drawbacks in static and dynamic quality. Instead, utilizing 3DGS to maintain an accurate head structure, our method simplifies the learning difficulty of facial motions with a pure deformation representation, therefore improving facial fidelity and lip-synchronization.

**Deformation in Radiance Fields.** Deformation has been widely applied in radiance fields to synthesize dynamic novel views. Some NeRF methods [13, 33, 34, 38, 41] use a static canonical radiance field to capture geometry and appearance and a time-dependent deformation field for dynamics. These methods predict an offset referring to the sampling position, which is opposite to the motion path and would bring extra difficulties in fitting. To solve this problem, [14] use a deformation that directly warps the canonical fields to represent dynamics. However, this method is costly since the spatial points cannot be accurately and stably controlled in its grid-based NeRF representation.

More recently, 3D Gaussian Splatting [20] introduces an explicit point-based representation for radiance fields, where deformation can be easily applied to a definite set of Gaussian primitives to directly warp the canonical fields. Based on this idea, considerable dynamic 3DGS works [22, 26, 30, 49, 51] get significant improvements in visual quality and efficiency for dynamic novel views synthesis. However, these methods only aim to remember the fixed motion at each time stamp, insufficient to represent various fine-grained motions driven by conditions, especially on the mouth. Despite some attempts conducted to reconstruct the human head [8, 39, 45, 50] driven by parametrized facial models, the mapping from audio to these parameters is not easy to learn and would cause information loss, and thus they can not be easily transferred to our audio-driven task. In this paper, we introduce deformable Gaussian fields with an incremental sampling strategy to facilitate learning multiple complex facial motions from a monocular speech video via pure deformation, and decomposite inconsistent motions of the face and inside mouth areas to improve the quality of delicate talking motions.

## 3 Method

### 3.1 Preliminaries and Problem Setting

**3D Gaussian Splatting.** 3D Gaussian splatting (3DGS) [20] represents 3D information with a set of 3D Gaussians. It computes pixel-wise color $\mathcal{C}$ with

a set of 3D Gaussian primitives $\theta$ and the camera model information at the observing view. Specifically, a Gaussian primitive can be described with a center $\mu \in \mathbb{R}^3$, a scaling factor $s \in \mathbb{R}^3$, and a rotation quaternion $q \in \mathbb{R}^4$. For rendering purposes, each Gaussian primitive also retains an opacity value $\alpha \in \mathbb{R}$ and a $Z$-dimensional color feature $f \in \mathbb{R}^Z$. Thus, the $i$-th Gaussian primitive $\mathcal{G}_i$ keeps a set of parameters $\theta_i = \{\mu_i, s_i, q_i, \alpha_i, f_i\}$. Its basis function is in the form of:

$$\mathcal{G}_i(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})^T \Sigma_i^{-1}(\mathbf{x}-\mu_\mathbf{i})}, \tag{1}$$

where the covariance matrix $\Sigma$ can be calculated from $s$ and $q$.

During the point-based rendering, a rasterizer would gather $N$ Gaussians following the camera model to compute the color $\mathcal{C}$ of pixel $\mathbf{x}_p$, with the decoded color $c$ of feature $f$ and the projected opacity $\widetilde{\alpha}$ calculated by their projected 2D Gaussians $\mathcal{G}^{proj}$ on image plane:

$$\mathcal{C}(\mathbf{x}_p) = \sum_{i \in N} c_i \widetilde{\alpha}_i \prod_{j=1}^{i-1}(1-\widetilde{\alpha}_j), \quad \widetilde{\alpha}_i = \alpha_i \mathcal{G}_i^{proj}(\mathbf{x}_p). \tag{2}$$

Similarly, the opacity $\mathcal{A} \in [0,1]$ of pixel $\mathbf{x}_p$ can be given:

$$\mathcal{A}(\mathbf{x}_p) = \sum_{i \in N} \widetilde{\alpha}_i \prod_{j=1}^{i-1}(1-\widetilde{\alpha}_j). \tag{3}$$

3DGS optimizes the parameters $\theta$ for all Gaussians through gradient descent under color supervision. During the optimization process, it applies a densification strategy to control the growth of the primitives, while also pruning unnecessary ones. This work inherits these optimization strategies for color supervision.

**Problem Setting.** In this paper, we aim to present an audio-driven framework based on 3DGS representation for high-fidelity talking head synthesis. Adopting a similar problem setting as NeRF-based works [15, 24, 27, 43], we take a few-minute speech video with a single person as the training data. A 3DMM model [35] is utilized to estimate the head pose and therefore to infer the camera pose. To keep aligned with previous works [15, 24, 27, 40], we use a pre-trained DeepSpeed [16] model as the basic audio encoder to get a generalizable audio feature from the raw input speech audio.

### 3.2 Deformable Gaussian Fields for Talking Head.

Despite previous NeRF-based methods [15, 24, 27, 40, 43, 52] have achieved great success in synthesizing high-quality talking heads via generating point-wise appearance, they can still not tackle the problem of generating distorted facial features on dynamic regions. One main reason is the appearance space, including color and density, is jumping and unsmooth, which makes it difficult for the continuous and smooth neural fields to fit. In comparison, deformation is another choice to represent motions with better smoothness and continuity, as shown in Fig. 3. In this work, we propose to purely use deformation in the Gaussian radiance fields to represent different motions of the talking head in 3D space. In particular, the whole representation is decomposed into Persistent Gaussian

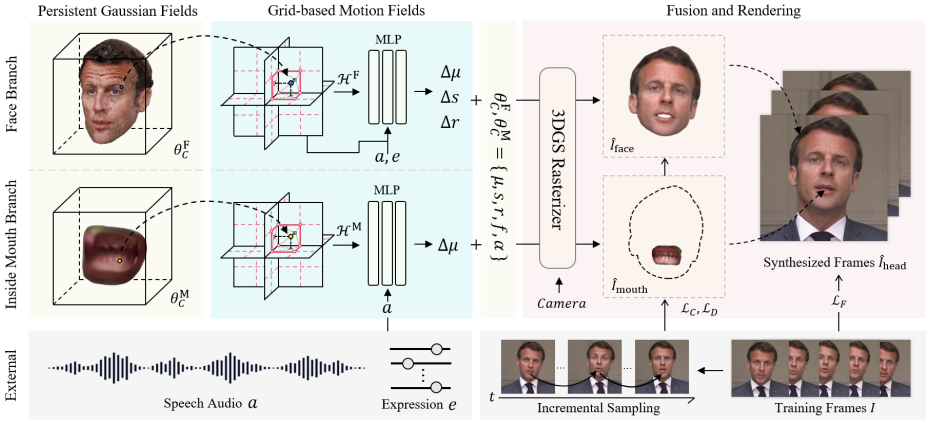


**Fig. 2: Overview of TalkingGaussian.** Learning from the speech video with training frames $I$, TalkingGaussian builds two separate branches to represent the dynamic face and inside mouth areas. Queried by the primitives in Persistent Gaussian Fields with parameters $\theta_C$, a point-wise deformation can be predicted from Grid-based Motion Fields conditioned with audio feature $\boldsymbol{a}$ and upper-face expression $\boldsymbol{e}$. After that, the 3DGS rasterizer renders the deformed 3D Gaussian primitives into 2D images observed from the given camera, which are then fused to synthesize the entire talking head.

Fields and Grid-based Motion Fields, as shown in Fig. 2. These fields will be further refined for different regions in the next section.

**Persistent Gaussian Fields.** Persistent Gaussian Fields preserve the persistent Gaussian primitive with the canonical parameters $\theta_C = \{\mu, s, q, \alpha, f\}$. Firstly, we initialize this module with the static 3DGS by the speech video frames to get a coarse mean field. Later, it attends a joint optimization with the Grid-based Motion Fields.

**Grid-based Motion Fields.** Although the primitives in Persistent Gaussian Fields can effectively represent the correct 3D head, a regional position encoding is lacking due to their fully explicit space structure. Considering most facial motions are regionally smooth and continuous, we adopt an efficient and expressive tri-plane hash encoder $\mathcal{H}$ [24] for position encoding with an MLP decoder to build Grid-based Motion Fields for a continuous deformation space.

Specifically, the motion fields aim to represent the facial motion by predicting a point-wise deformation $\delta_i = \{\Delta\mu_i, \Delta s_i, \Delta q_i\}$ for each primitive with the input of its center $\mu_i$, which is irrelevant to the color and opacity changing. For the given condition feature set $\mathbf{C}$, the deformation $\delta_i$ can be calculated by:

$$\delta_i = \text{MLP}(\mathcal{H}(\mu_i) \oplus \mathbf{C}), \tag{4}$$

where $\oplus$ denotes concatenation.

Through a 3DGS rasterizer, these two fields are combined to generate deformed Gaussian primitives to render the output image, of which the deformed

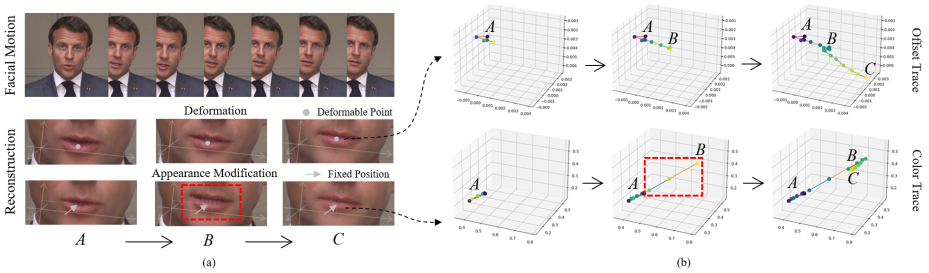


**Fig. 3: (a)** The reconstructed facial motion results represented by deformation and appearance modification. **(b)** The visualized traces of the changing coordinate offset (deformation) and color in RGB (appearance modification) of two points with the same initial position. During the process, offset changes smoothly and the corresponding results are clear and accurate. Instead, some sudden changes with a large step length may occur in color, which is difficult to fit and causes a distorted mouth (red box).

parameters $\theta_D$ are got from the canonical parameters $\theta_C$ and deformation $\delta$:

$$\theta_D = \{\mu + \Delta\mu, s + \Delta s, q + \Delta q, \alpha, f\}. \tag{5}$$

**Optimization with Incremental Sampling.** While learning the deformation, once the target primitive position is too far from the predicted results, the gradient would vanish and thus the motion fields may fail to be effectively updated. To tackle this problem, we introduce an incremental sampling strategy. Specifically, we first find a valid metric $m$ (e.g. action units [11] or landmarks) to measure the deformation degree of each target facial motion. Then, at the $k$-th training iteration, we use a sliding window to sample a required training frame at position $j$, of which the motion metric $m_j$ satisfies the condition:

$$m_j \in [B_{lower} + k \times T, B_{upper} + k \times T], \tag{6}$$

where $B_{lower}$ and $B_{upper}$ denote the initial lower and upper bound of the sliding window, and $T$ denotes the step length. This selected training frame can offer sufficient new knowledge for the deformable fields to learn, but would not be too hard. To avoid catastrophic forgetting, we apply the incremental sampling strategy once every $K$ iterations.

### 3.3 Face-Mouth Decomposition

Although the Grid-based Motion Fields can predict the point-wise deformation at arbitrary positions due to the continuous and dense 3D space representation, this representation still encounters a granularity problem caused by the motion inconsistency between the face and the inside mouth. Since the inside area of the mouth is spatially too close to the lips but does not always move together, their motions would interfere with each other in a single interpolation-based motion field. This can also further lead to a bad reconstruction quality in static structure as well, as shown in Fig. 4.

To tackle this problem, we propose decomposing these two regions in 3D space and building two individual branches with separate optimization. For each

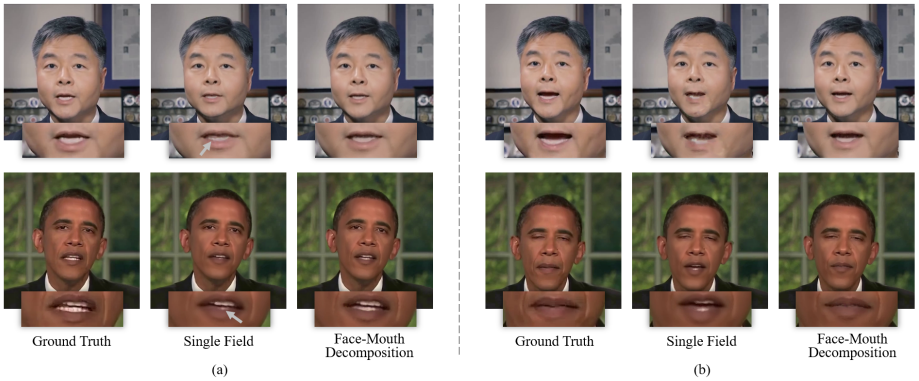


**Fig. 4: (a)** Lips and the inside mouth, especially teeth, are hard to be correctly divided with a single motion field. **(b)** This would further affect the learning of the mouth structure and speaking motions, resulting in bad quality. Our Face-Mouth Decomposition can successfully address this problem and render high-fidelity results.

training video frame, we first use the off-the-shelf face parsing models to get a semantic mask of the inside mouth region in 2D space [6]. Then, we take the masked image of the inside mouth and the remaining surface region (containing the face, hair, and other head parts) to train two separate deformable Gaussian fields as two branches of our framework.

**Face Branch.** The face branch serves as the main part to fit the appearance and motion of the talking head, including all facial motions except the one of the inside mouth. In this branch, we adopt a region attention mechanism [24] in the Grid-based Motion Fields to facilitate the learning of the conditioned deformation driven by the features of audio $\boldsymbol{a}$ and upper-face expression $\boldsymbol{e}$. To fully decouple these two conditions, the upper-face expression feature $\boldsymbol{e}$ is composed of 7 action units [11] that are explicitly irrelevant to the mouth. The deformation $\delta_i^{\mathrm{F}}$ for the $i$-th primitive in the face branch can be predicted by:

$$\delta_i^{\mathrm{F}} = \mathrm{MLP}(\mathcal{H}^{\mathrm{F}}(\mu_i) \oplus \boldsymbol{a}_{r,i} \oplus \boldsymbol{e}_{r,i}), \tag{7}$$

where $\boldsymbol{a}_{r,i} = V_{\boldsymbol{a},i} \odot \boldsymbol{a}$ and $\boldsymbol{e}_{r,i} = V_{\boldsymbol{e},i} \odot \boldsymbol{e}$ denote the region-aware feature at position $\mu_i$ in the region attention mechanism, calculated by the attention vectors $V_{\boldsymbol{a},i}$ and $V_{\boldsymbol{e},i}$ with the Hadamard product $\odot$.

During the optimization, we apply the Incremental Sampling strategy for the lips action and eye-blinking. Specifically, we measure the lips opening degree by the height of the mouth area according to the detected facial landmarks, and use AU45 [11] to describe the degree of eye close. Then, we gradually move the sliding window to guide the face branch to learn the deformations of the lips from close to open and eyes from open to close.

**Inside Mouth Branch.** The inside mouth branch represents the audio-driven dynamic inside mouth region in 3D space. Considering the inside mouth moves in a much simpler manner and is only driven by audio, we use a lightweight

[6] Additional descriptions and details can be found in the supplementary material.

deformable Gaussian field to build this branch. In particular, we only predict the translation $\Delta\mu_i$ conditioned by the audio feature $\boldsymbol{a}$ for the $i$-th primitive:

$$\delta_i^{\mathrm{M}} = \{\Delta\mu_i^{\mathrm{M}}\} = \mathrm{MLP}(\mathcal{H}^{\mathrm{M}}(\mu_i) \oplus \boldsymbol{a}). \tag{8}$$

To get a better reconstruction quality of the teeth part, we apply an incremental sampling strategy that smooths the learning of the overlapping between teeth and lips with the quantitative metric AU25 [11].

**Rendering.** The final talking head image is fused with the two rendered face and inside mouth images. Based on the physical structure, we assume the rendering results from the Inside Mouth Branch are behind that from the Face Branch. Therefore, the talking head color $\mathcal{C}_{\mathrm{head}}$ of pixel $\mathbf{x}_p$ can be rendered by:

$$\mathcal{C}_{\mathrm{head}}(\mathbf{x}_p) = \mathcal{C}_{\mathrm{face}}(\mathbf{x}_p) \times \mathcal{A}_{\mathrm{face}}(\mathbf{x}_p) + \mathcal{C}_{\mathrm{mouth}}(\mathbf{x}_p) \times (1 - \mathcal{A}_{\mathrm{face}}(\mathbf{x}_p)), \tag{9}$$

where $\mathcal{C}_{\mathrm{face}}$ and $\mathcal{A}_{\mathrm{face}}$ denote the predicted face color and opacity from the face branch, and $\mathcal{C}_{\mathrm{mouth}}$ is the color predicted by the inside mouth branch.

### 3.4 Training Details

We keep the basic 3DGS optimization strategies to train our framework. The full process can be divided into three stages, of which the first two stages are individually applied for the two branches and the last stage is for fusion.

**Static Initialization.** At the beginning of the training, we first conduct an initialization via the vanilla 3DGS for the Persistent Gaussian Fields to get a coarse head structure. Following 3DGS, we use a pixel-wise L1 loss and a D-SSIM term to measure the error between the image $\hat{\mathcal{I}}_C$ rendered by parameters $\theta_C$ and the masked ground-truth image $\mathcal{I}_{\mathrm{mask}}$ for each branch:

$$\mathcal{L}_C = \mathcal{L}_1(\hat{\mathcal{I}}_C, \mathcal{I}_{\mathrm{mask}}) + \lambda\mathcal{L}_{\mathrm{D-SSIM}}(\hat{\mathcal{I}}_C, \mathcal{I}_{\mathrm{mask}}). \tag{10}$$

**Motion Learning.** After the initialization, we add the motion fields into training via its predicted deformation $\delta$. In practice, we take the deformed parameters $\theta_D$ from Equation 5 as the input for the 3DGS rasterizer to render the output image $\hat{\mathcal{I}}_D$. The loss function is:

$$\mathcal{L}_D = \mathcal{L}_1(\hat{\mathcal{I}}_D, \mathcal{I}_{\mathrm{mask}}) + \lambda\mathcal{L}_{\mathrm{D-SSIM}}(\hat{\mathcal{I}}_D, \mathcal{I}_{\mathrm{mask}}). \tag{11}$$

**Fine-tuning.** Finally, a color fine-tuning stage is conducted to better fuse the head and inside mouth branches. We calculate the reconstruction loss between the fused image $\hat{\mathcal{I}}_{\mathrm{head}}$ rendered by Equation 9 and the ground-truth video frame $\hat{\mathcal{I}}$ with pixel-wise L1 loss, D-SSIM, and LPIPS terms:

$$\mathcal{L}_F = \mathcal{L}_1(\hat{\mathcal{I}}_{\mathrm{head}}, \mathcal{I}) + \lambda\mathcal{L}_{\mathrm{D-SSIM}}(\hat{\mathcal{I}}_{\mathrm{head}}, \mathcal{I}) + \gamma\mathcal{L}_{\mathrm{LPIPS}}(\hat{\mathcal{I}}_{\mathrm{head}}, \mathcal{I}). \tag{12}$$

At this stage, we only update the color parameter $f \in \theta_C$ and stop the densification strategy of 3DGS for stability.

# 4 Experiment

## 4.1 Experimental Settings

**Dataset.** We collect four high-definition speaking video clips from previous publicly-released video sets [15, 24, 52] to build the video datasets for our experiments, including three male portraits "Macron", "Lieu", "Obama", and one female portrait "May". The video clips have an average length of about 6500 frames in 25 FPS with a center portrait, three ("May", "Macron", and "Lieu") of which are cropped and resized to $512 \times 512$ and one ("Obama") to $450 \times 450$.

**Comparison Baselines.** In the experiments, we mainly compare our method with the most related NeRF-based methods AD-NeRF [15], DFRF [40], RAD-NeRF [43], GeneFace [52] and ER-NeRF [24], which render talking head via person-specific radiance fields trained with speech videos. Additionally, we also take the state-of-the-art 2D generative models (Wav2Lip [37], IP-LAP [58] and DINet [57]), which do not need person-specific training, and person-specific methods (SynObama [42], NVP [44], and LSP [29]) as the baselines.

**Implementation Details.** Our method is implemented on PyTorch. For a specific portrait, we first train both the face and inside mouth branches for $50,000$ iterations parallelly and then jointly fine-tune them for $10,000$ iterations. Adam [21] and AdamW [28] optimizers are used in training. In the loss functions, $\lambda$ and $\gamma$ are set to 0.2 and 0.5. All experiments are performed on RTX 3080 Ti GPUs. The overall training process takes about 0.5 hours. A pre-trained DeepSpeech model [16] is used as a basic audio feature extractor.

## 4.2 Quantitative Evaluation

**Comparison Settings.** To evaluate the reconstruction quality and lip-audio synchronization ability, our quantitative comparison contains two settings: **1)** The *self-reconstruction setting*, where we split each of the four videos into training and test sets, and use the audio, expression, and pose sequences in the unseen test set to reconstruct the talking head in a self-driven way for quality evaluation. **2)** The *lip-synchronization setting*, where we use the audio track from other videos to drive the models trained in the first setting and evaluate lip-synchronization, focusing on situations with cross-domain input audios. Specifically, we use the same audio samples as previous works [24, 36] from NVP and SynObama as two test audios A and B to evaluate the "Obama" and "May" portraits. Since both audio A and B are from unseen videos with male voices, the evaluation results, especially on "May" with a different gender, can well illustrate the generalization ability. Tests for NVP, SynObama, and SSP-NeRF are conducted only on their released demos due to the lack of training codes.

**Metrics and Measurements.** In the aspect of static image quality, we employ PSNR for the overall quality, LPIPS [56] for high-frequency details, and SSIM [47] to evaluate face structure. For dynamic motions, we also utilize the landmark

**Table 1: The quantitative results of the *self-reconstruction setting*.** The best and second-best methods are in **bold** and underline, respectively.

| Methods | Rendering Quality | | | Motion Quality | | | Efficiency | |
|---|---|---|---|---|---|---|---|---|
| | PSNR ↑ | LPIPS ↓ | SSIM ↑ | LMD ↓ | AUE-(L/U) ↓ | Sync-C ↑ | Time | FPS |
| Ground Truth | N/A | 0 | 1.000 | 0 | 0/0 | 7.584 | - | - |
| Wav2Lip [37] | - | - | - | 6.861 | 1.46 / - | **8.749** | - | 21.6 |
| IP-LAP [58] | **35.34** | 0.0405 | 0.903 | 5.601 | 0.77 / - | 4.897 | - | 3.18 |
| DINet [57] | 32.08 | 0.0393 | 0.856 | 6.411 | 0.97 / - | 6.321 | - | 27.2 |
| AD-NeRF [15] | 31.87 | 0.0942 | 0.877 | 2.791 | 0.71/1.26 | 5.353 | 18.7h | 0.11 |
| DFRF [40] | 31.73 | 0.0858 | 0.876 | 3.406 | 0.74/1.40 | 4.127 | 22.4h | 0.04 |
| RAD-NeRF [43] | 33.07 | 0.0530 | 0.887 | 2.761 | 0.65/1.14 | 5.052 | 5.3h | 28.7 |
| GeneFace [52] | 30.49 | 0.0670 | 0.846 | 3.339 | 1.28/1.34 | 5.291 | 5.8h | 20.9 |
| ER-NeRF [24] | 32.83 | 0.0289 | 0.889 | 2.676 | 0.55/0.88 | 5.295 | 2.1h | 31.2 |
| ER-NeRF+$\boldsymbol{e}$ | 33.14 | 0.0271 | 0.902 | 2.623 | 0.57/0.31 | 5.754 | - | - |
| **Ours** | 33.61 | **0.0259** | **0.910** | **2.586** | **0.53/0.22** | 6.516 | **0.5h** | **108** |

**Table 2: The quantitative results of the *lip-synchronization setting*.** The best and second-best methods are in **bold** and underline, respectively.

| Method | Test Audio A | | | | Test Audio B | | | |
|---|---|---|---|---|---|---|---|---|
| | "Obama" | | "May" | | "Obama" | | "May" | |
| | Sync-E ↓ | Sync-C ↑ | Sync-E ↓ | Sync-C ↑ | Sync-E ↓ | Sync-C ↑ | Sync-E ↓ | Sync-C ↑ |
| Ground Truth | 0 | 6.701 | 0 | 6.701 | 0 | 7.309 | 0 | 7.309 |
| LSP [29] | 8.683 | 5.045 | 9.511 | 4.441 | 8.640 | 5.504 | 9.882 | 4.167 |
| SynObama [42] | 8.197 | **6.802** | - | - | - | - | - | - |
| NVP [44] | - | - | - | - | 10.175 | 4.316 | - | - |
| AD-NeRF [15] | 9.742 | 5.195 | 9.517 | 4.757 | 10.682 | 4.314 | 9.518 | 5.319 |
| DFRF [40] | 10.662 | 3.905 | 10.830 | 3.135 | 11.044 | 3.690 | 11.248 | 3.215 |
| RAD-NeRF [43] | 9.552 | 5.585 | 11.883 | 2.000 | 8.680 | 6.667 | 11.176 | 2.426 |
| GeneFace [52] | 9.052 | 5.336 | 10.259 | 3.569 | 8.966 | 5.674 | 10.173 | 4.280 |
| ER-NeRF [24] | 9.123 | 6.134 | 10.251 | 3.639 | 8.688 | 6.706 | 10.535 | 4.141 |
| ER-NeRF+$\boldsymbol{e}$ | 9.573 | 6.092 | 9.825 | 4.012 | 8.934 | 6.577 | 11.226 | 4.423 |
| **Ours** | **8.635** | 5.962 | **9.368** | **4.774** | **8.627** | **6.737** | **9.273** | **5.441** |

distance (LMD) [6] and the confidence score (Sync-C) and error distance (Sync-D) of SyncNet [9,10] for lip synchronization. Additionally, we estimate the action units [11] of the videos by OpenFace [2,3] and divide them into an upper-face action unit error (AUE-U) and lower-face action unit error (AUE-L) according to their definitions to separately evaluate the upper-face and mouth motions.

In the first setting, we measure the PSNR and LPIPS on the whole image, and SSIM on the face region. We also record the person-specific training time and inference FPS of all methods. Since all the 2D-based generative baselines do not modify the upper part of the face, we do not measure their AUE-U. Notably, we provide another video clip as the image reference for Wav2Lip to avoid information leakage, for which PSNR, LPIPS, and SSIM are not valid. In the second setting, we use the non-comparison-based Sync-C and Sync-E to quantitatively measure the lip-synchronization quality.

**Evaluation Results.** We report the results of the two settings in Table 1 and Table 2, respectively. Considering the upper-face expression condition would also influence performance [24,36], we add our upper-face feature $\boldsymbol{e}$ to ER-NeRF [24] as "ER-NeRF+$\boldsymbol{e}$" for a fair comparison. **1)** In the *self-reconstruction setting*, our method achieves the best overall image quality, motion quality, and efficiency. For image quality, our method performs best in rendering accurate details (LPIPS)


**Fig. 5: Qualitative comparison of visual-audio synchronization.** Our method performs best in synthesizing accurately synchronized talking head compared with all baselines [15, 24, 37, 40, 43, 52, 57, 58]. Please **zoom in for better visualization.**

and structure (SSIM), due to our deformation-based motion representation and persistent head structure. In the aspect of motion quality, our method outperforms all NeRF methods in all metrics. Notably, TalkingGaussian gets a Sync-C score even higher than the generative method IP-LAP and DINet, demonstrating the powerful modeling ability of our method. Although Wav2Lip gets the best scores in Sync-C, its shortcomings in preserving personal talking styles lead to poor AUE-L and LMD. Moreover, due to the efficiency improvement brought by 3DGS, our method reaches the fastest training and inference speed in all baselines. **2)** In the results in the *lip-synchronization setting*, our method shows the best generalization performance. Although ER-NeRF can also get good scores on "Obama", it performs much worse in the more challenging cross-gender situation of "May". This phenomenon can also be observed in many other NeRF-based methods, especially RAD-NeRF which keeps a complex audio encoding module. Surprisingly, AD-NeRF performs well in this situation, despite having a relatively blurry rendering. Besides the difference in model quality, we consider this decreasing generalizability also to be caused by the overfitting of the audio feature when previous NeRF-based methods try to fit the unsmooth changing appearance to reconstruct delicate talking heads. Instead, when just using the same unimproved audio feature extractor as most previous baselines [15, 24, 40, 44], our method can simultaneously keep high-level static rendering quality and best generalization ability for various training videos and input audios, thanks to our simpler and smoother deformation-based motion representation.

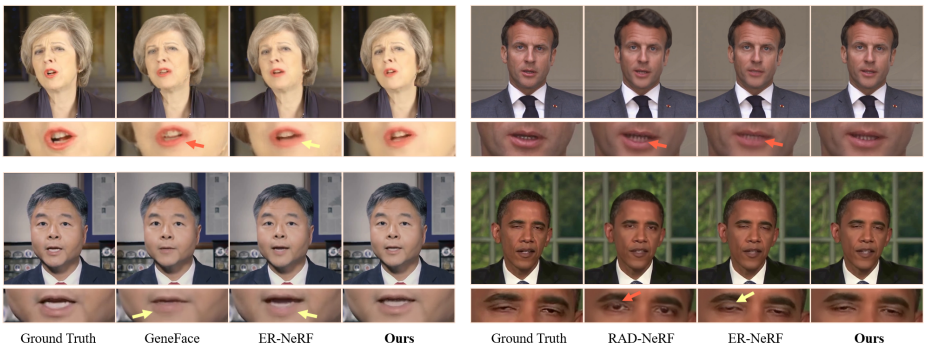


**Fig. 6: Qualitative comparison of the generated facial details.** Our method synthesizes more accurate and intact details than the recent NeRF-based state-of-the-art methods [24, 43, 52]. Please **zoom in for better visualization.**

### 4.3 Qualitative Evaluation

**Evaluation Results.** To qualitatively evaluate the synthesis quality, we show the keyframes of a reconstructed sequence from the *self-reconstruction setting* and details of four portraits in Fig. 5 and 6. **1)** Comparing the synthesized motion sequence in Fig. 5, our TalkingGaussian outperforms other methods by generating better visual-audio synchronized results. The generative methods (Wav2Lip, IP-LAP, and DINet) are short in generating high-quality images, as the trade-off for their one or few-shot ability. Lack of precise control signals, most NeRF-based baselines can not control audio-independent actions like eye blinking (orange arrow). While most NeRF-based baselines fail to synthesize some difficult mouth actions (blue box), our TalkingGaussian can precisely reappear these motions, without introducing any advanced audio encoders [36, 52]. This demonstrates the effectiveness of the learning task simplifications brought by our two decomposition designs. **2)** The comparison of synthesized details in Fig. 6 shows our advantages in facial fidelity and fidelity. As illustrated in Sec. 3.2, both RAD-NeRF and ER-NeRF exhibit distorted (red arrow) and blurry (yellow arrow) facial features in the dynamic regions. Even using the more precise facial landmarks to condition the NeRF renderer, GeneFace still can't escape from this inherent trouble brought by the previous appearance-modification paradigm. By purely representing motions with deformation, our method tackles this problem and succeeds in synthesizing more accurate and intact facial features.

**User Study.** To better judge the visual quality in real scenarios judged by humans, we conducted a user study where a total of 32 talking head videos were generated by 8 methods. Then we invited 16 attendees to rate these methods according to their generated results from three aspects: (1) Lip-sync Accuracy; (2) Video Realness; and (3) Image Quality. The results are reported in Table 3, in which our TalkingGaussian performs the best in all three aspects, demonstrating the potential value of our method in real-world applications.

**Table 3: User Study.** The rating is in the range of 1-5, higher denotes better. We highlight the **best** and second best results.

| Methods | Wav2Lip [37] | IP-LAP [58] | DI-Net [57] | AD-NeRF [15] | GeneFace [52] | RAD-NeRF [43] | ER-NeRF [24] | **TalkingGaussian** |
|---|---|---|---|---|---|---|---|---|
| Lip-sync Accuracy | 2.50 | 1.63 | 3.25 | 2.75 | 3.13 | 3.19 | 3.56 | **3.94** |
| Image Quality | 1.75 | 2.44 | 2.69 | 3.25 | 3.69 | 3.31 | 3.63 | **4.06** |
| Video Realness | 1.69 | 1.88 | 1.88 | 3.19 | 3.31 | 3.19 | 3.44 | **3.88** |

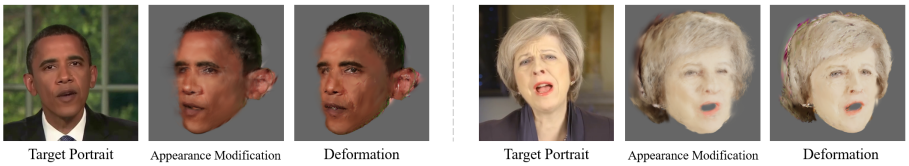


**Fig. 7: 3D visualization** of the heads generated by deformation and appearance modification on 3DGS. Deformation performs better in generating more precise geometry.

### 4.4 Ablation Study

To prove the effectiveness of our contributions, we conduct the ablation study using the self-reconstruction setting. The results are reported in Table 4.

**Motion Representation.** First, we use the same module as our motion fields to predict the deformation and appearance modification for the Tri-Hash backbone from ER-NeRF [24] and 3DGS [20] to evaluate the motion representations. Due to the lack of accurate point-wise controls, deformation performs worse on the NeRF-based Tri-Hash. After introducing 3DGS, deformation shows its advantage in preserving better facial features and bringing higher image quality scores. We also visualize the results on 3DGS with the same conditions in Fig. 7 for comparison. The results demonstrate the effectiveness of both our 3DGS-based persistent head structure and deformation-based motion representation.

**Face-Mouth Decomposition** (FMD). We apply our FMD to the 3DGS backbone to illustrate its effect. Although the results show that FMD can help in lip-synchronization in all situations, the improvement is much larger when combined with deformation, since it can solve the face-mouth motion conflict that seriously affects deformation learning. With this combination, our framework successfully reaches the best motion quality, especially in lip-synchronization.

**Incremental Sampling** (IS). Raised by the unstable optimization process of deformation, some jitters, incomplete motions, and errors in geometry structure can be observed in the generated videos, which lead to a lower SSIM. These problems are relieved after applying IS, demonstrating its contribution to generating more smooth and realistic talking heads.

## 5 Conclusion

This paper represents a novel deformation-based framework TalkingGaussian for high-quality 3D talking head synthesis. Our study is the first to reveal a "facial distortion" problem caused by inaccurate predictions of the rapidly changing

**Table 4: Ablation Study** of our contributions under the self-reconstruction setting.

| Backbone | Appearance | Deformation | FMD | IS | PSNR ↑ | LPIPS ↓ | SSIM ↑ | LMD ↓ | AUE-(L/U) ↓ | Sync-C ↑ |
|---|---|---|---|---|---|---|---|---|---|---|
| Tri-Hash | ✓ | | | | 33.14 | 0.0271 | 0.902 | 2.623 | 0.57/0.31 | 5.754 |
| | | ✓ | | ✓ | 31.50 | 0.0334 | 0.877 | 3.016 | 0.67/0.38 | 5.285 |
| 3DGS | ✓ | | | | 33.34 | 0.0355 | 0.904 | 2.630 | 0.56/0.25 | 6.001 |
| | | ✓ | | ✓ | 33.42 | 0.0290 | 0.903 | 2.665 | 0.54/0.23 | 5.676 |
| | ✓ | | ✓ | | 33.27 | 0.0351 | 0.904 | 2.605 | 0.55/0.24 | 6.332 |
| | | ✓ | ✓ | | 33.57 | 0.0260 | 0.906 | **2.584** | **0.53**/0.23 | 6.497 |
| | | ✓ | ✓ | ✓ | **33.61** | **0.0259** | **0.910** | 2.586 | **0.53**/**0.22** | **6.516** |

appearance. By maintaining a persistent head structure with 3DGS and decomposing inconsistent motions into different spaces, TalkingGaussian addresses this problem with a deformation paradigm and achieves superior performance in synthesizing realistic and accurate talking heads compared to existing methods.

**Ethical Consideration.** We hope our method can promote the healthy development of digital industries. However, it must be noted that our method may be misused for malicious purposes and cause negative influence. We recommend the responsible use of this technique. As part of our responsibility, we will also assist in developing deepfake detection techniques by sharing our generated results.

# Supplementary Material for TalkingGaussian

Jiahe Li[1], Jiawei Zhang[1], Xiao Bai[1⋆], Jin Zheng[1], Xin Ning[2], Jun Zhou[3], and Lin Gu[4,5]

[1] School of Computer Science and Engineering, State Key Laboratory of Complex & Critical Software Environment, Jiangxi Research Institute, Beihang University
[2] Institute of Semiconductors, Chinese Academy of Sciences
[3] School of Information and Communication Technology, Griffith University
[4] RIKEN AIP
[5] The University of Tokyo

## Overview

In the supplementary material, we first report additional experiments in Sec. A. We further show an additional visualization in Sec. B, and discuss the responsibility to human subjects and ethical consideration in Sec. C and D. Limitations and future work are summarized in Sec. E. A supplementary video is also provided as an additional illustration.

## A Additional Experiments

### A.1 Hybrid Motion Representation

We additionally conduct an experiment to explore whether a hybrid representation of mixing deformation and appearance modification could benefit the performance. The results are reported in Table 5. Upon the basic deformation $\delta$, we first evaluate the setting of additionally predicting a factor to adjust the opacity $\alpha$. Then we further try to predict the RGB color of each primitive directly rather than using the SH feature $f$. The results show that: 1) Adding *alpha* would not help more. This is reasonable since the opacity of each primitive should be a static value. 2) Adding RGB would result in a blurry rendering. In the first aspect, it would bring a larger burden for the network to store more information. On the other hand, a non-persistent color can still suffer from the distortion problem from inaccurate appearance prediction to some extent.

### A.2 Extension

Besides the experiment settings in the main paper, our method is scalable and can extend to a wider range of applications.

⋆ Corresponding author: Xiao Bai (baixiao@buaa.edu.cn).

**Table 5:** Exploration of hybrid motion representations.

| Settings | PSNR ↑ | LPIPS ↓ | SSIM ↑ |
|---|---|---|---|
| $\delta + \alpha$ | 33.60 | 0.0261 | 0.908 |
| $\delta + \alpha +$ RGB | **33.63** | 0.0264 | 0.907 |
| $\delta$ | 33.61 | **0.0259** | **0.910** |

**Table 6:** Exploration of adopting different audio encoders under *self-reconstruction setting*. The best and second-best results are in **bold** and underline.

| Extractor | Rendering Quality | | | Motion Quality | | |
|---|---|---|---|---|---|---|
| | PSNR ↑ | LPIPS ↓ | SSIM ↑ | LMD ↓ | AUE-(L/U) ↓ | Sync-C ↑ |
| Ground Truth | N/A | 0 | 1.000 | 0 | 0/0 | 7.584 |
| DeepSpeech | **33.61** | 0.0259 | 0.910 | 2.586 | 0.53/**0.22** | 6.516 |
| Wav2Vec 2.0 | 33.59 | 0.0260 | **0.911** | **2.582** | **0.52**/0.23 | 6.552 |
| HuBERT | 33.60 | **0.0258** | 0.909 | 2.583 | **0.52**/0.24 | **6.667** |

**Audio Feature Extractor.** Following previous baselines, we have adopted a pre-trained DeepSpeech [16] model to extract audio features in the main experiments, for a fair comparison. In fact, our method can also easily connect to more powerful feature extractors and become stronger. Table 6 shows our performance with different audio feature extractors Wav2Vec 2.0 [1] and HuBERT [17] under the *self-reconstruction setting*. The results demonstrate a high-quality audio feature could boost the performance of TalkingGaussian, especially on lip-synchronization, showing the growth potential of our approach.

**Cross-Lingual and Cross-Gender.** Our method can also applied to more challenging cross-lingual and cross-gender cases. In this experiment, we collect 4 training videos, in which 2 males and 2 females with English and French audio are included, and use challenging 3 test audio clips to drive their corresponding models. The three test audios consist of two German audio clips separately with a female and a male voice, and a Chinese audio clip with a male voice. In this setting, we compare our method to two baselines ER-NeRF [24] and GeneFace [52] that utilize different extractors.

In Table 7, the results show that our models all perform better than the baselines while using the same extractors. Notably, GeneFace has further used the HuBERT features to pre-train an intermediate representation on a large audio-visual corpus to enhance its generalizability for cross-domain audios. The generated results have been provided in the supplementary video.

**Singing.** While inputting a song, our method can even synthesize high-quality singing talking heads with no such training audio included. This demonstrates our surprising generalization ability and robustness for cross-domain inputs, and shows applicability for a wider range of situations. The generated videos can be found in the supplementary video.

**Table 7:** Exploration of cross-lingual and cross-gender situations. The best results for each audio feature extractor are in **bold**.

| Extractor | Methods | *Female, German* | | *Male, German* | | *Male, Chinese* | |
|---|---|---|---|---|---|---|---|
| | | Sync-E ↓ | Sync-C ↑ | Sync-E ↓ | Sync-C ↑ | Sync-E ↓ | Sync-C ↑ |
| DeepSpeech | ER-NeRF | 9.773 | 3.273 | 10.497 | 3.381 | 10.577 | 2.893 |
| | **Ours** | **9.399** | **3.720** | **9.677** | **4.407** | **10.322** | **3.378** |
| HuBERT | GeneFace | 8.753 | 4.059 | 9.208 | 4.969 | 10.597 | 3.720 |
| | **Ours** | **8.260** | **4.691** | **8.323** | **5.556** | **8.856** | **4.539** |

## B Additional Visualization

Here we show some high-definition synthesized frames in Fig. 8 for a convenient and intuitive visualization. Compared with current SOTA methods GeneFace [52] and ER-NeRF [24], our method performs best in image quality while retaining a high lip-sync accuracy. We have also provided more additional results in our supplementary video. We strongly recommend watching it for better visualization.

## C Dataset Declaration

In the experiments, all of the multimedia datasets we used were obtained from existing works [15, 24, 52]. To our knowledge, most of these data are collected from the internet. In our work, we have tried our best to use data containing only public figures to avoid invading personal privacy. All the data are manually checked to reduce the existence of offensive content.

## D Ethical Consideration

As our target, we hope our TalkingGaussian can promote the healthy development of digital industries. However, it must be noted that our method may be misused for malicious purposes and cause negative influence. We recommend the responsible use of this technique:

- **Informed Consent.** Whenever this technique is in use for spread purposes, ensure that all individuals in the training data have provided explicit, informed consent.
- **Disclosure.** Please disclose the use of our method, and any other deepfake techniques as well, in all synthesized products. This is critical to ensure all audiences are aware that the content is real, and may include misleading information.

For protection purposes, we will support the development of more powerful deepfake detectors to alert people to the presence of fake content.

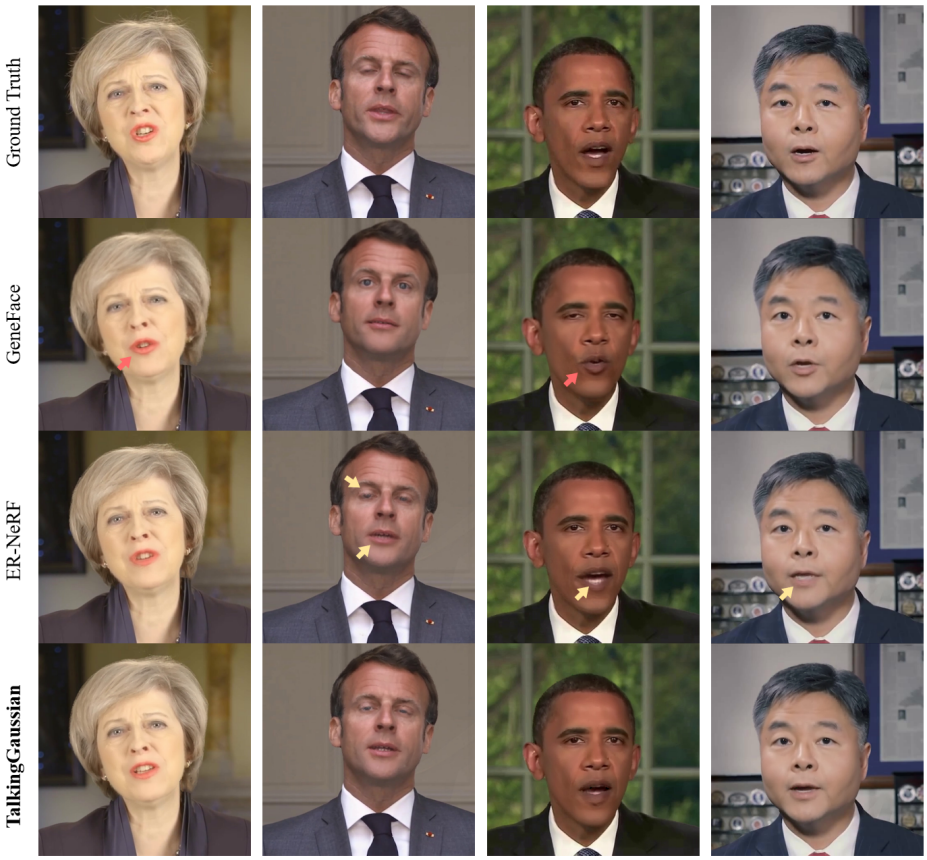


**Fig. 8: Additional High-definition Comparisons.** ER-NeRF [24] heavily sufferers the facial distortion problem caused by inaccurate appearance prediction. GeneFace [52] performs better in preserving fidelity, since it has introduced an intermediate representation to bridge the audio-visual mapping. However, its synchronization quality drops. In comparison, our method synthesizes better talking heads both in static and dynamic.

## E Limitations and Future Work

In this paper, our proposed TalkingGaussian outperforms in rendering high-quality lip-synchronized talking head videos, with better facial fidelity and higher efficiency than previous methods. Despite that, our method still has some limitations.

In the first aspect, some noisy primitives may randomly occur due to the densification operation of 3DGS. Although this can be relieved by a smoother optimization process provided by Incremental Sampling, it would still sometimes influence the quality. In future works, we will consider adding more constraints to better control the primitive's growth.

On the other hand, the face and inside mouth branches are aligned via the audio feature in our method, enabling free and individual learning for their own motions. Nevertheless, this connection is not tight enough. Although it is sufficient for most in-domain audio inputs like speeches from the same person as that in the training data, the face and inside mouth area may be misaligned in some cross-domain situations. To solve this problem, we may build a better awareness of these two parts to enhance robustness in the future.

## References


1. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)
2. Baltrušaitis, T., Mahmoud, M., Robinson, P.: 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). vol. 6, pp. 1–6. IEEE (2015)
3. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). pp. 59–66. IEEE (2018)
4. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16123–16133 (2022)
5. Chatziagapi, A., Athar, S., Jain, A., Rohith, M., Bhat, V., Samaras, D.: Lipnerf: What is the right feature space to lip-sync a nerf? In: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–8. IEEE (2023)
6. Chen, L., Li, Z., Maddox, R.K., Duan, Z., Xu, C.: Lip movements generation at a glance. In: Computer Vision–ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII 15. pp. 538–553. Springer (2018)
7. Chen, L., Maddox, R.K., Duan, Z., Xu, C.: Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7832–7841 (2019)
8. Chen, Y., Wang, L., Li, Q., Xiao, H., Zhang, S., Yao, H., Liu, Y.: Monogaussianavatar: Monocular gaussian point-based head avatar. arXiv preprint arXiv:2312.04558 (2023)
9. Chung, J.S., Zisserman, A.: 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. pp. 87–103. Springer (2017)
10. Chung, J.S., Zisserman, A.: Out of time: Automated lip sync in the wild. In: Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13. pp. 251–263. Springer (2017)
11. Ekman, P., Friesen, W.V.: Facial Action Coding System: Manual. Palo Alto: Consulting Psychologists Press (1978)
12. Ezzat, T., Geiger, G., Poggio, T.: Trainable videorealistic speech animation. ACM Transactions on Graphics (TOG) **21**(3), 388–398 (2002)
13. Fang, J., Yi, T., Wang, X., Xie, L., Zhang, X., Liu, W., Nießner, M., Tian, Q.: Fast dynamic radiance fields with time-aware neural voxels. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022)
14. Guo, X., Sun, J., Dai, Y., Chen, G., Ye, X., Tan, X., Ding, E., Zhang, Y., Wang, J.: Forward flow for novel view synthesis of dynamic scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16022–16033 (2023)
15. Guo, Y., Chen, K., Liang, S., Liu, Y.J., Bao, H., Zhang, J.: Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5784–5794 (2021)

16. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al.: Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567 (2014)
17. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: 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)
18. Jamaludin, A., Chung, J.S., Zisserman, A.: You said that?: Synthesising talking faces from audio. International Journal of Computer Vision **127**, 1767–1779 (2019)
19. Kapitanov, A., Kvanchiani, K., Sofia, K.: Easyportrait - face parsing and portrait segmentation dataset. arXiv preprint arXiv:2304.13509 (2023)
20. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
22. Kratimenos, A., Lei, J., Daniilidis, K.: Dynmf: Neural motion factorization for real-time dynamic view synthesis with 3d gaussian splatting. arXiv preprint arXiv:2312.00112 (2023)
23. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
24. Li, J., Zhang, J., Bai, X., Zhou, J., Gu, L.: Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7568–7578 (2023)
25. Li, W., Zhang, L., Wang, D., Zhao, B., Wang, Z., Chen, M., Zhang, B., Wang, Z., Bo, L., Li, X.: One-shot high-fidelity talking-head synthesis with deformable neural radiance field. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17969–17978 (2023)
26. Lin, Y., Dai, Z., Zhu, S., Yao, Y.: Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. arXiv preprint arXiv:2312.03431 (2023)
27. Liu, X., Xu, Y., Wu, Q., Zhou, H., Wu, W., Zhou, B.: Semantic-aware implicit neural audio-driven video portrait generation. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII. pp. 106–125. Springer (2022)
28. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
29. Lu, Y., Chai, J., Cao, X.: Live speech portraits: real-time photorealistic talking-head animation. ACM Transactions on Graphics (TOG) **40**(6), 1–17 (2021)
30. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. arXiv preprint arXiv:2308.09713 (2023)
31. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision. pp. 405–421. Springer (2020)
32. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022)
33. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5865–5874 (2021)

34. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. arXiv preprint arXiv:2106.13228 (2021)
35. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: 2009 sixth IEEE international conference on advanced video and signal based surveillance. pp. 296–301. Ieee (2009)
36. Peng, Z., Hu, W., Shi, Y., Zhu, X., Zhang, X., He, J., Liu, H., Fan, Z.: Synctalk: The devil is in the synchronization for talking head synthesis. arXiv preprint arXiv:2311.17590 (2023)
37. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: 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. pp. 484–492 (2020)
38. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10318–10327 (2021)
39. Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Giebenhain, S., Nießner, M.: Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. arXiv preprint arXiv:2312.02069 (2023)
40. Shen, S., Li, W., Zhu, Z., Duan, Y., Zhou, J., Lu, J.: Learning dynamic facial radiance fields for few-shot talking head synthesis. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII. pp. 666–682. Springer (2022)
41. Song, L., Chen, A., Li, Z., Chen, Z., Chen, L., Yuan, J., Xu, Y., Geiger, A.: Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. arXiv preprint arXiv:2210.15947 (2022)
42. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics (ToG) **36**(4), 1–13 (2017)
43. Tang, J., Wang, K., Zhou, H., Chen, X., He, D., Hu, T., Liu, J., Zeng, G., Wang, J.: Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. arXiv preprint arXiv:2211.12368 (2022)
44. Thies, J., Elgharib, M., Tewari, A., Theobalt, C., Nießner, M.: 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. pp. 716–731. Springer (2020)
45. Wang, J., Xie, J.C., Li, X., Xu, F., Pun, C.M., Gao, H.: Gaussianhead: High-fidelity head avatars with learnable gaussian derivation (2024)
46. Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., Loy, C.C.: Mead: A large-scale audio-visual dataset for emotional talking-face generation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI. pp. 700–717. Springer (2020)
47. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
48. Wiles, O., Koepke, A.S., Zisserman, A.: X2face: A network for controlling face generation using images, audio, and pose codes. In: Computer Vision–ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII 15. pp. 690–706. Springer (2018)

49. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. arXiv preprint arXiv:2310.08528 (2023)
50. Xu, Y., Chen, B., Li, Z., Zhang, H., Wang, L., Zheng, Z., Liu, Y.: Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. arXiv preprint arXiv:2312.03029 (2023)
51. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. arXiv preprint arXiv:2309.13101 (2023)
52. Ye, Z., Jiang, Z., Ren, Y., Liu, J., He, J., Zhao, Z.: Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. In: The Eleventh International Conference on Learning Representations (2022)
53. Ye, Z., Zhong, T., Ren, Y., Yang, J., Li, W., Huang, J., Jiang, Z., He, J., Huang, R., Liu, J., et al.: Real3d-portrait: One-shot realistic 3d talking portrait synthesis. arXiv preprint arXiv:2401.08503 (2024)
54. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 325–341 (2018)
55. Zhang, C., Zhao, Y., Huang, Y., Zeng, M., Ni, S., Budagavi, M., Guo, X.: Facial: Synthesizing dynamic talking face with implicit attribute learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3867–3876 (2021)
56. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
57. Zhang, Z., Hu, Z., Deng, W., Fan, C., Lv, T., Ding, Y.: Dinet: Deformation inpainting network for realistic face visually dubbing on high resolution video. arXiv preprint arXiv:2303.03988 (2023)
58. Zhong, W., Fang, C., Cai, Y., Wei, P., Zhao, G., Lin, L., Li, G.: Identity-preserving talking face generation with landmark and appearance priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2023)