GaussianSpeech: Audio-Driven Personalized 3D Gaussian Avatars

Shivangi Aneja[1] Artem Sevastopolsky[1] Tobias Kirschstein[1] Justus Thies[2,3]
Angela Dai[1] Matthias Nießner[1]
[1]Technical University of Munich [2]MPI-IS, Tübingen [3]TU Darmstadt

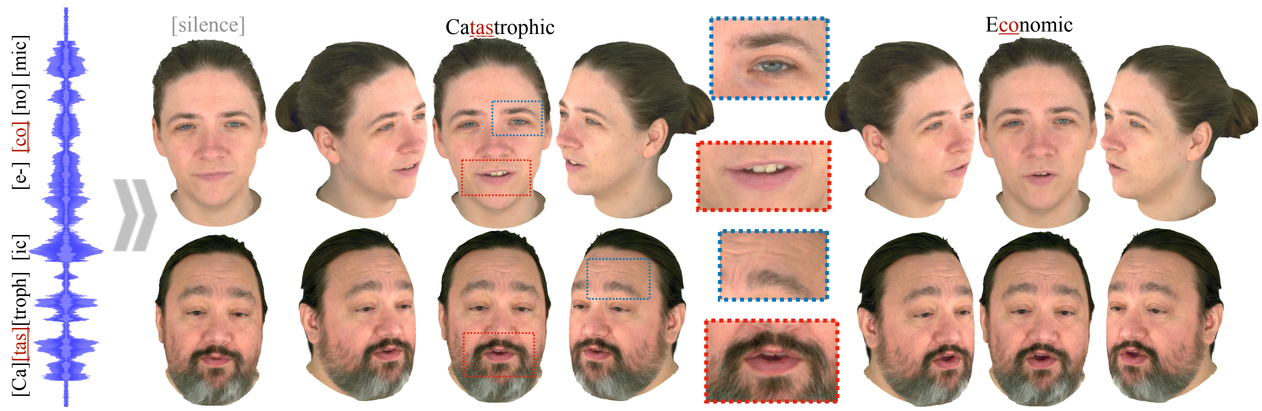

Figure 1. Given input speech signal, GaussianSpeech can synthesize photorealistic 3D-consistent talking human head avatars. Our method can generate realistic and high-quality animations, including mouth interiors such as teeth, wrinkles, and specularities in the eyes. We handle diverse facial geometry, including hair buns and mustaches/beards, while effectively synchronizing to the audio signal.

Abstract

We introduce GaussianSpeech[1], a novel approach that synthesizes high-fidelity animation sequences of photo-realistic, personalized 3D human head avatars from spoken audio. To capture the expressive, detailed nature of human heads, including skin furrowing and finer-scale facial movements, we propose to couple speech signal with 3D Gaussian splatting to create realistic, temporally coherent motion sequences. We propose a compact and efficient 3DGS-based avatar representation that generates expression-dependent color and leverages wrinkle- and perceptually-based losses to synthesize facial details. To enable sequence modeling of 3D Gaussian splats with audio, we devise an audio-conditioned transformer model capable of extracting lip and expression features directly from audio input. Due to the absence of high-quality dataset of talking humans in correspondence with audio, we captured a new large-scale multi-view dataset of audio-visual sequences of talking humans with native English accents and diverse facial geometry. GaussianSpeech consistently achieves state-of-the-art quality with visually natural motion, while encompassing diverse facial expressions and styles.

[1]Project Page: https://shivangi-aneja.github.io/projects/gaussianspeech

1. Introduction

Generating animated sequences of photorealistic 3D head avatars from spoken audio is important for many graphics applications, including immersive telepresence, movies, and virtual assistants. In particular, rendering photorealistic views of such animated avatars from various viewpoints is crucial for realistic, immersive digital media, for instance, telepresence to a meeting room requires a photorealistic appearance for all viewpoints of the people in the room, or AR/VR where users can freely change their viewpoint.

Creating such photorealistic animated 3D avatars from audio remains challenging, as it requires maintaining photorealistic fidelity throughout the animation sequence, as well as from various viewpoints. Existing work thus focuses on addressing these objectives independently; various works focus on re-enacting videos in the 2D domain [3, 8, 18, 29, 30, 37, 42, 44, 65, 66, 68], creating front-view video animations, while others focus on animating 3D face geometry from audio [15, 40, 51, 63]. In contrast, we aim to create innately 3D audio-driven avatars enabling 3D-consistent, free-viewpoint photorealistic synthesis needed for immersive digital communication.

In order to characterize audio-driven 3D animation of a person from multi-view input, we propose to represent animated head sequences with explicit 3D Gaussian points, leveraging the detailed and expressive representation space of 3D Gaussian Splatting (3DGS) [27]. 3DGS offers a flexible representation capable of handling complex and irregular facial geometry and appearance (e.g., different skin tones, beard, skin creasing) and real-time rendering, making it a well-suited choice for facial animation.

Thus, we design an efficient, personalized 3D Gaussian avatar representation from multi-view input observations of a person, containing relatively few Gaussian splats in order to make sequence modeling of photorealistic 3DGS tractable and allowing us to operate at real-time rendering rates. This is achieved through learning expression- and view-dependent color, and our losses focusing on perceptual face quality using a face recognition network, as well as focusing on fine-scale details through wrinkle detection.

Our efficient, high-quality avatar can handle the nuances of the facial geometry, like skin tone variation and dynamic wrinkles. We then use this person-specific avatar to guide audio-driven head animation, enabled by our transformer-based sequence model. We learn lip motion features and wrinkle features directly from audio to obtain expression input to train our transformer model, enabling photorealistic generation of a coherent animation sequence.

To create high-fidelity, audio-driven animated 3D head avatars, we require high-resolution multi-view data paired with high-quality audio recordings. Existing multiview datasets [28, 58] unfortunately lack either high-quality video or high-quality audio captures. In the absence of large-scale and high-quality paired audio-multiview data of people speaking, we collected a new multiview dataset with 16 cameras for 6 native English participants captured at 30 fps and 3208x2200 resolution with overall recordings of $\sim$3.5 hours, an order of magnitude larger than the existing datasets. We will make the dataset and the corresponding 3D face trackings publicly available for research purposes. To summarize, this paper makes the following contributions:

- The first transformer-based sequence model for audio-driven head animation synthesis of a lightweight 3DGS based avatar. By animating our optimized 3DGS avatar directly with our transformer model, we achieve temporally coherent animation sequences while characterizing fine-scale face details and speaker-specific style.
- A new high-quality audio-video dataset, comprising high-resolution 16-view dataset of 6 native English speakers (Standard American & British). The dataset has a total of 2500 sequences, with overall recordings of $\sim$3.5 hours.

2. Related Work

Audio-driven facial animation plays an important role in digital media. Here we discuss audio-driven animation methods generating different output representations.

2.1. 2D-Based Methods.

There is a large corpus of works in the field of 2D audio-driven facial animation operating on monocular RGB videos, synthesizing 2D sequences directly [4–7, 9, 12, 17, 19, 20, 23, 25, 35, 38, 43, 46, 47, 53, 55–57, 59, 60, 62, 64, 69, 70]. However, these methods operate in pixel space and can produce very limited side views. Another line of work also operating on frontal RGB videos but using intermediate 3D representations are based on 3DMMs [14, 22, 45, 48, 52, 67]. Although these methods generate photorealistic results, they use 3DMMs as a proxy to improve the animation quality and are still limited to frontal and limited side views. In contrast, we model head avatars with explicit 3D Gaussian points, thus, enabling simultaneous free-viewpoint rendering for different viewpoints which is critical for telepresence applications.

2.2. Parametric Model Based Methods.

Another promising line of work is to animate 3D facial geometry directly. A vast majority of these works model speech-conditioned animation for either artist-designed template meshes [10, 11, 15, 26, 40, 50, 51, 63] or blend-shapes for 3D parametric head model [1, 36]. While these methods can faithfully match facial motion with the speech signal and can be rendered from different viewpoints, they do not model any appearance or texture information and cannot handle complex and irregular facial geometry. The synthesized animations, therefore, do not look realistic. Compared to these, our method optimizes a 3DGS-based avatar and models appearance using expression and view-dependent color, generating photorealistic results.

2.3. Radiance Fields Based Methods.

Recent speech-driven animation methods based on radiance fields [18, 29, 32, 37, 42, 65, 66] have gained popularity due to their ability to model directly from images. Neural Radiance Fields (NeRF) [34] possess the capability to render a scene from arbitrary viewpoints, however, existing audio-driven methods utilizing NeRF are designed for monocular videos. Concurrent to ours, few recent works [8, 21, 30] leverage 3DGS [27] for generating audio-driven talking heads. GaussianTalker [8] and Talking-Gaussian [30] focus on improving the rendering speed for monocular videos. EmoTalk3D [21] can synthesize multi-view renders, however these methods generate sequences frame-by-frame, thus suffer from jitter and scaling artefacts. In contrast, our method synthesizes multi-view consistent and temporally smooth results, including fine-scale details like dynamic wrinkles, by leveraging a transformer-based sequence model and an efficient 3DGS-based avatar.

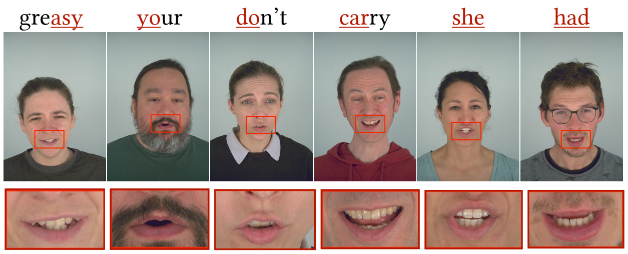

Figure 2. Random frames selected for each participant (top) from the dataset and corresponding zoom-in for the mouth region (bottom). We captured a gender-balanced dataset of native speakers with different English accents and diverse facial geometry including different skin tones, beard and glasses to maximize diversity.

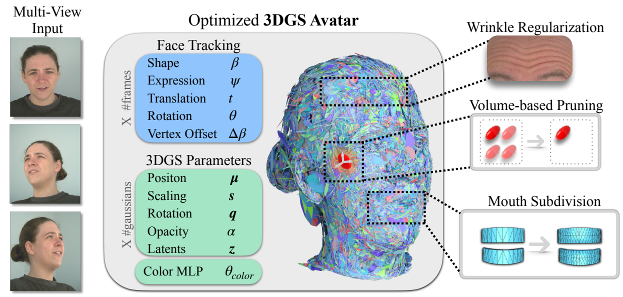

Figure 3. Person-specific 3D Avatar: We compute 3D face tracking and bind 3D Gaussians to the triangles of the tracked FLAME mesh. We apply volume-based pruning to prevent optimization to generate large amount of Gaussians, and apply subdivision of mesh triangles in the mouth region. We train color MLP θ_{color} to synthesize expression & view dependent color. We apply wrinkle regularization and perceptual losses to improve photorealism.

3. Multi-View Audio-Visual Dataset

We collected a novel dataset consisting of six native English speakers captured using a multiview rig of 16 cameras (see Supp.). We record sequences at 30 FPS at 3208 x 2200 resolution. To achieve quality and diversity, we specifically capture native English speakers with different accents, including American, British, and Canadian. We selected participants aged 20-50 with different genders and facial geometry including beard and glasses to increase the diversity, see Fig. 2. We collected 415 sequences for every subject, leading to an overall recording time of 30-35 minutes for each of the 16 cameras. The spoken sentences are chosen from the TIMIT [16] corpus to maximize the phonetic diversity. Our dataset stands out from the existing datasets in terms of quality and quantity.

While certain datasets with audio-visual talking faces exist, they are limited in quality. The RAVDESS dataset [33] contains a set of native speakers, but it has only 2 unique sequences per participant with North American accent, while we captured three different English accents and 415 unique sentences. The MEAD dataset [58] captured the participants with 250 unique sentences per participant. However, they focus on emotional speech synthesis due to which they capture only 40 unique natural expression/emotion per participant at a relatively lower resolution. The Nersemble [28] dataset captures the participants at high resolution, but it only contains 10 audio sequences per participant. Closest to ours is MultiFace [61], which captured participants in a spherical rig of 150 cameras; however, it captured only 50 audio sequences per participant. Our dataset contains 415 sequences for every subject at high resolution, an order of magnitude larger than existing datasets, see Tab. 1. We plan to release our entire dataset to the research community.

4. Method

Our method operates in two stages. First, we develop a lightweight and high-quality avatar initialization based on GaussianAvatars (Sec. 4.1). Next, we train a transformer-based sequence model to animate our initialized avatar conditioned on personalized audio features (Sec. 4.2). Since our method requires 3D face tracking, we compute them from our multiview sequence dataset, similar to [39].

Dataset	# Cam	# Unique Sentences	Resolution	Duration (in minutes/camera)	Native
RAVDESS [33]	1	2	1920 x 1080	0.1 min	✓
MEAD [58]	8	250	1920 x 1080	20 min	✗
EmoTalk3D [21]	11	N/A	512 x 512	20 min	✗
Nersemble [28]	16	10	**3208** x **2200**	1 min	✗
MultiFace [61]	**150**	50	2048 x 1334	4 min	✗
Ours	16	**415**	**3208** x **2200**	**35 min**	✓

Table 1. Existing Audio-Video Dataset Comparison per participant in the datasets. Compared to existing datasets, ours is an order of magnitude larger and higher resolution. All datasets are captured at standard 30 fps.

4.1. Avatar Initialization

We propose an efficient optimization strategy to compute a 3DGS-based Gaussian avatar representation. We found that naively training GaussianAvatar [39] generates blurred/low-quality textures, especially, for scenarios with rapid facial movement like faster talking speed/head motion. In addition, GaussianAvatar can not effectively handle dynamic wrinkles. Therefore, we introduce expression-dependent colors and propose several regularizations to improve quality of our avatars described below and shown in Fig. 3.

Volume-Based Pruning. We modify the pruning strategy used by GaussianAvatar. Instead of pruning 3D Gaussian splats based on a given opacity threshold $\epsilon_{\text{opacity}}$, we select top 25,000 Gaussians with maximum opacity and 3D Gaussian's scale volume combined at every pruning step as $\mathcal{G}_i = \sigma_i \cdot (s_x \cdot s_y \cdot s_z)$, where σ_i refers to i^{th} Gaussian's opacity and s_x, s_y, s_z refers to its scale along x, y, and z axis. Even when the optimization generates excessive splats during densification, this top-k pruning ensures that the optimized avatar does not contain too many 3D Gaussian splats. However, this leads to degradation in quality by removing small transparent 3D splats and generates blurry results. We, thus, propose to add additional regularizations to improve quality.

Expression-dependent Color. Instead of learning SH Color for 3D Gaussians, our method generates color with a lightweight two-layer color MLP $\boldsymbol{\theta}_{\text{color}}$ to faithfully synthesize dynamic wrinkles. Given a FLAME [31] expression code $\boldsymbol{\psi}$ and viewing direction $\boldsymbol{v}$, we synthesize view- and expression-dependent color $\boldsymbol{c}_i$ as $\boldsymbol{c}_i = \boldsymbol{\theta}_{\text{color}}(\boldsymbol{\psi}; \boldsymbol{z}_i; \boldsymbol{v})$. Note that we additionally learn per Gaussian latent features $\boldsymbol{z}_i$ for sharper colors.

Perceptual Losses. To improve the sharpness of the color generated by $\boldsymbol{\theta}_{\text{color}}$, we add a global and patch-based perceptual loss. The global perceptual loss $\mathcal{L}_{\text{global}}$ is based on the content and style features of the pre-trained face recognition model ArcFace [13]. The content loss $\mathcal{L}_{\text{content}}$ and style loss $\mathcal{L}_{\text{style}}$ are defined as:

$$\mathcal{L}_{\text{content}} = \sum_{k=1}^{K} \left|\left|\phi_k(I_{\text{render}}) - \phi_k(I_{gt})\right|\right|_1, \quad (1)$$

$$\mathcal{L}_{\text{style}} = \sum_{k=1}^{K} \left|\left|\mathcal{G}_k(I_{\text{render}}) - \mathcal{G}_k(I_{gt})\right|\right|_1, \quad (2)$$

where ϕ_k and $\mathcal{G}_k$ refer to the feature maps and Gram matrices [24] for the layer k respectively. I_{render} and I_{gt} refer to the rendered and ground-truth multiview image.

$$\mathcal{L}_{\text{global}} = \mathcal{L}_{\text{content}} + \mathcal{L}_{\text{style}}. \quad (3)$$

$\mathcal{L}_{\text{global}}$ improves the quality of the texture globally, however, it shows limited improvements for fine-scale skin areas and less observed regions like the mouth interior. We, therefore, employ a VGG-based loss on local image patches based on content features of the pre-trained VGG backbone:

$$\mathcal{L}_{\text{patch}} = \frac{1}{J} \sum_{j=1}^{J} \sum_{k=1}^{K} \left|\left|\zeta_k(I^j_{\text{render}}) - \zeta_k(I^j_{\text{gt}})\right|\right|_1, \quad (4)$$

where I^j_{render} and I^j_{gt} refer to the j^{th} local patch regions from the rendered and ground-truth multiview images. We use 128×128 patches and sample 16 local patches uniformly for the facial area by employing alpha matting.

Wrinkle Regularization. Naive optimization of GaussianAvatar [39] cannot represent skin creasing and fine-scale wrinkles, since it learns a constant color for the avatar, irrespective of facial expression. To overcome this, we introduce a lightweight color MLP θ_{color} that can generate expression-dependent wrinkles. We employ a novel wrinkle feature loss $\mathcal{L}_{\text{wrinkle}}$ which focuses on refining dynamic wrinkles. Specifically, we run an off-the-shelf wrinkle detector [41] to extract wrinkle features and apply a content loss on its feature detection backbone during optimization:

$$\mathcal{L}_{\text{wrinkle}} = \sum_{k=1}^{K} \left|\left|\Psi_k(I_{\text{render}}) - \Psi_k(I_{\text{gt}})\right|\right|_1. \quad (5)$$

Note that our method synthesizes wrinkles faithfully for avatars whose captured data includes dynamic wrinkles when speaking; if the avatar did not display wrinkles during speech, our method will not generate them.

Mouth Region Subdivision. Since the mouth interior (especially teeth) is less frequently observed compared to other facial regions, the standard 3DGS-based densification cannot generate sufficient Gaussians for the mouth to synthesize high quality results. To address this, before optimization, we subdivide the triangles which are used to initialize the Gaussians corresponding to the teeth in the FLAME mesh using a uniform four-way subdivision. By doing so, we begin with a high density of Gaussians for the teeth, compensating for low gradient magnitude in this area, ensuring that teeth appear detailed and realistic.

To summarize, we optimize our 3DGS-based avatar using $\mathcal{L}_{\text{total}}$ loss as:

$$\begin{aligned}\mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{rgb}} + \lambda_{\text{pos}}\mathcal{L}_{\text{position}} + \lambda_{\text{s}}\mathcal{L}_{\text{scaling}} \\ &+ \lambda_{\text{g}}\mathcal{L}_{\text{global}} + \lambda_{\text{p}}\mathcal{L}_{\text{patch}} + \lambda_{\text{w}}\mathcal{L}_{\text{wrinkle}},\end{aligned} \quad (6)$$

where $\mathcal{L}_{\text{rgb}}, \mathcal{L}_{\text{position}}, \mathcal{L}_{\text{scaling}}$ are defined in [39] (also explained in Supp. doc).

4.2. Sequence Model Training

GaussianSpeech performs high-fidelity and temporally-consistent generative synthesis of avatar motion sequences, conditioned on audio signal. To characterize complex face motions and fine-scale movements like dynamic wrinkles, we employ a transformer-based sequence model. We predict mesh animations with our sequence model and refine the dynamic motion attributes of the 3D Gaussian Splats of our optimized avatar to be consistent with audio features. An overview of our approach is illustrated in Fig. 4.

Audio Encoding. We employ the state-of-the-art pre-trained speech model Wav2Vec 2.0 [2] to encode the audio signal. Specifically, we use the audio feature extractor made up of temporal convolution layers (TCN) to extract audio feature vectors $\{a_i\}_{i=1}^{N_a}$ from the raw waveform, followed by a *Frequency Interpolation* layer to align the input audio signal $\{a_i\}_{i=1}^{N_a}$ (captured at frequency f_a = 16kHz) with our dataset $\{a_i\}_{i=1}^{N_e}$ (captures at framerate f_e = 30FPS).

Lip Features. A stacked multi-layer *Lip Transformer Encoder* processes these resampled audio features and predicts personalized lip content feature vectors $\boldsymbol{c}^{1:T}$. To avoid learning spurious correlation between upper face motion and audio, the Lip Transformer Encoder is trained with only lip vertices from the FLAME mesh with L2-reconstruction loss autoregressively as:

$$\mathcal{L}_{\text{lip}} = \sum_{n=1}^{N} \left(\sum_{t=1}^{T} \left|\left|\boldsymbol{l}^t_{\text{gt}} - \boldsymbol{l}^t_{\text{pred}}\right|\right|_2 \right)_n, \quad (7)$$

where T refers to the number of frames per sequence and N total sequences, $\boldsymbol{l}_{\text{gt}}$ and $\boldsymbol{l}_{\text{pred}}$ refer to the ground truth and predicted lip vertices, respectively.

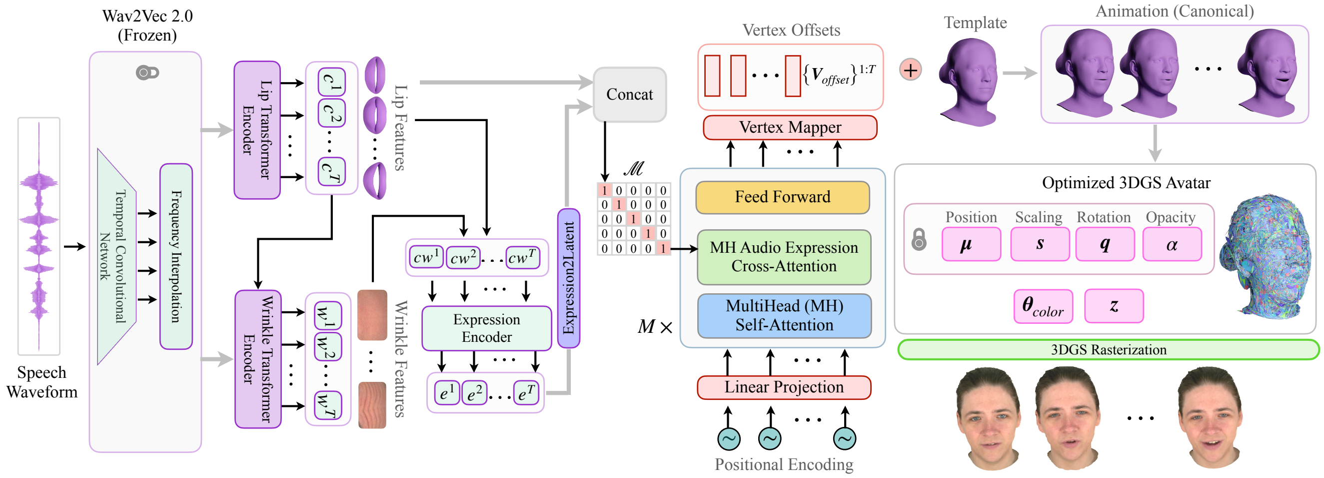

Figure 4. Method Overview. From the given speech signal, GaussianSpeech uses Wav2Vec 2.0 [2] encoder to extract generic audio features and maps them to personalized lip feature embeddings $\boldsymbol{c}^{1:T}$ with Lip Transformer Encoder and wrinkle features $\boldsymbol{w}^{1:T}$ with Wrinkle Transformer Encoder. Next, the Expression Encoder synthesizes FLAME expressions $\boldsymbol{e}^{1:T}$ which are then projected via Expression2Latent MLP and concatenated with $\boldsymbol{c}^{1:T}$ for input to the motion decoder. The motion decoder employs a multi-head transformer decoder [54] consisting of Multihead Self-Attention, Cross-Attention, and Feed Forward layers. The concatenated lip-expression features are fused into the decoder via cross-attention layers with alignment mask $\mathcal{M}$. The decoder then predicts FLAME vertex offsets $\{\boldsymbol{V}_{\text{offset}}\}^{1:T}$ which gets added to the template mesh $\boldsymbol{T}$ to generate vertex animation in canonical space. During training, these are then fed to our optimized 3DGS avatar (Sec. 4.1) and the color MLP $\boldsymbol{\theta}_{\text{color}}$ and gaussian latents $\boldsymbol{z}$ are further refined via re-rendering losses [27].

Wrinkle Features. Similarly, our *Wrinkle Transformer Encoder* conditioned on audio and lip features predicts personalized wrinkle feature vectors $\boldsymbol{w}^{1:T}$. The Wrinkle Transformer Encoder is trained with wrinkle features extracted using a wrinkle detector [41] from the RGB frames as:

$$\mathcal{L}_{\text{wrinkle}} = \sum_{n=1}^{N}\left(\sum_{t=1}^{T}||\boldsymbol{w}_{\text{gt}}^{t} - \boldsymbol{w}_{\text{pred}}^{t}||_2\right)_n, \tag{8}$$

where $\boldsymbol{w}_{\text{gt}}$ and $\boldsymbol{w}_{\text{pred}}$ refer to the ground truth and predicted wrinkle vertices respectively.

Expression Features. Using personalized lip features $\boldsymbol{c}^{1:T}$ and wrinkle features $\boldsymbol{w}^{1:T}$ obtained above, we train the *Expression Encoder* $\mathcal{E}_{\text{exp}}$. Specifically, we concatenate lip and wrinkle features to obtain combined features $\boldsymbol{cw}^{1:T} = \left[\boldsymbol{c}^{1:T}; \boldsymbol{w}^{1:T}\right]$. These combined features are fed to our Expression Encoder which predicts FLAME expressions as $\boldsymbol{e}_{\text{pred}}^{1:T} = \mathcal{E}_{\text{exp}}(\boldsymbol{cw}^{1:T})$ and is trained with:

$$\mathcal{L}_{\text{expr}} = \sum_{n=1}^{N}\left(\sum_{t=1}^{T}||\boldsymbol{e}_{\text{gt}}^{t} - \boldsymbol{e}_{\text{pred}}^{t}||_2\right)_n, \tag{9}$$

where $\boldsymbol{e}_{\text{gt}}$ and $\boldsymbol{e}_{\text{pred}}$ refers to the ground truth and predicted FLAME expression parameters, respectively.

Audio-Conditioned Animation. We train a transformer decoder [54] network to synthesize mesh *Vertex Offsets* $\left\{\boldsymbol{V}_{\text{offset}}\right\}^{1:T}$, where T refers to the number of frames in a sequence. During training, we first project the predicted expression parameters $\boldsymbol{e}_{\text{pred}}^{1:T}$ via the *Expression2Latent MLP* $\mathcal{E}$ to the latent space of our model and concatenate it with lip features $\boldsymbol{c}^{1:T}$ to obtain combined lip-expression motion features $\boldsymbol{m}^{1:T} = \left[\boldsymbol{c}^{1:T}; \mathcal{E}(\boldsymbol{e}^{1:T})\right]$.

These motion features $\boldsymbol{m}^{1:T}$ are then processed through transformer decoder, and the *Vertex Mapper MLP* to synthesize Vertex Offsets $\left\{\boldsymbol{V}_{\text{offset}}\right\}^{1:T}$ in canonical space. We leverage a look-ahead binary target mask $\mathcal{T} \in \mathbb{R}^{N\times N}$ in the multi-head self-attention layer to prevent the model from peeking into the future frames. The $(i,j)^{th}$ element of the matrix with $1 \le i, j \le N$ is:

$$\mathcal{T}_{ij} = \begin{cases} True & \text{if } i \le j \\ False & \text{else} \end{cases} \tag{10}$$

Input motion features $\boldsymbol{m}^{1:T}$ are fused into the transformer with the multi-head audio expression cross-attention layer via the alignment mask $\mathcal{M}$. The binary mask $\mathcal{M} \in \mathbb{R}^{N\times N}$ is a Kronecker delta function δ_{ij} such that the motion features for i^{th} timestamp attend to vertex features at the j^{th} timestamp if and only if $i = j$:

$$\mathcal{M} = \delta_{ij} = \begin{cases} True & \text{if } i = j \\ False & \text{if } i \neq j \end{cases} \tag{11}$$

The vertex offsets are obtained as:

$$\left\{\boldsymbol{V}_{\text{offset}}\right\}^{1:T} = \mathcal{D}\Big(\boldsymbol{m}^{1:T} \,|\, \mathcal{T}, \mathcal{M}\Big), \tag{12}$$

where $\mathcal{D}$ refers to the transformer decoder network. These predicted offsets $\left\{\boldsymbol{V}_{\text{offset}}\right\}^{1:T}$ are added to the template mesh $\boldsymbol{T}$ to obtain mesh animation in canonical space as $\left\{\boldsymbol{V}_{\text{pred}}\right\}^{1:T} = \boldsymbol{T} + \left\{\boldsymbol{V}_{\text{offset}}\right\}^{1:T}$.

The *Expression2Latent MLP* $\mathcal{E}$ and the transformer decoder $\mathcal{D}$ are jointly trained with an L2-reconstruction loss:

$$\mathcal{L}_{\text{vertices}} = \sum_{n=1}^{N} \Big(\sum_{t=1}^{T} ||\boldsymbol{V}_{\text{gt}}^{t} - \boldsymbol{V}_{\text{pred}}^{t}||_2 \Big)_n, \tag{13}$$

The predicted vertices $\{\boldsymbol{V}_{\text{pred}}\}^{1:T}$ are fed to our Optimized 3DGS avatar (Sec. 4.1) and color related attributes of the avatar are further refined. We propose an alternating training strategy for the task as explained below.

(a) In the first step, we predict vertex displacements (from the rest pose) in the canonical space for the entire sequence (Eq. 12). This learns the optimal parameters for transformer $\mathcal{D}$ and Expression2Latent MLP $\mathcal{E}$ as:

$$\mathcal{E}^*, \mathcal{D}^* = \arg\min_{\mathcal{E},\mathcal{D}} \mathcal{L}_{\text{vertices}} \tag{14}$$

(b) In the second step, we predict the 3D Gaussian attributes with our Optimized 3DGS avatar (Sec. 4.1) and render the full animation sequence.

The color MLP $\boldsymbol{\theta}_{\text{color}}$ of our optimized avatar is conditioned on predicted FLAME expression $\boldsymbol{e}_{\text{pred}}$ and per Gaussian latent $\boldsymbol{z}_i$, in addition to view direction $\boldsymbol{v}$, and predicts the view- and expression-dependent color as $\boldsymbol{c}_i = \boldsymbol{\theta}_{\text{color}}(\boldsymbol{e}_{\text{pred}}; \boldsymbol{z}_i; \boldsymbol{v})$. The predicted image I_{pred} is obtained with the differentiable renderer $\mathcal{R}$ from Kerbl *et al.* [27] as:

$$I_{\text{pred}} = \mathcal{R}\big(\{\boldsymbol{\mu_i}, \boldsymbol{s_i}, \boldsymbol{q_i}, \boldsymbol{c_i}\}^{1:G}, [R\,|\,t]\big), \tag{15}$$

where $\boldsymbol{\mu_i}, \boldsymbol{s_i}, \boldsymbol{q_i}$ refers to the optimized avatar's position, scale, and rotations, respectively, and G defines the total number of Gaussians. The predictions are supervised with the photometric loss $\mathcal{L}_{\text{photo}}$ for the sequence:

$$\mathcal{L}_{\text{photo}} = \sum_{t=1}^{T} (\mathcal{L}_{\text{rgb}} + \lambda_{\text{g}}\mathcal{L}_{\text{global}} + \lambda_{\text{p}}\mathcal{L}_{\text{patch}})_t\,, \tag{16}$$

In this step, we refine per-Gaussian latents $\boldsymbol{z}_i$ and Color MLP $\boldsymbol{\theta}_{\text{color}}$ with audio-conditioned expressions:

$$\boldsymbol{\theta}_{\text{color}}^*, \{\boldsymbol{z}_i^*\}^{1:G} = \arg\min_{\boldsymbol{\theta}_{\text{color}}, \{\boldsymbol{z}_i\}^{1:G}} \mathcal{L}_{\text{photo}} \tag{17}$$

Overall, we optimize two losses in the alternating fashion: (a) $\mathcal{L}_{\text{vertices}}$ which learns audio-conditioned facial motion and (b) $\mathcal{L}_{\text{photo}}$ which refines the optimized avatar for more accurate and photorealistic appearance. We do not refine the position, scale, rotation, and opacity; empirically, we found that they did not make a noticeable difference in the overall quality.

5. Results

We evaluate GaussianSpeech on the tasks of (a) Avatar Representation and (b) Audio-Driven Animation. For (a), we evaluate standard perceptual image quality metrics SSIM, PSNR and LPIPS. For audio-driven animation, we evaluate lip synchronization LSE-D [38] as well as perceptual quality metrics. We train personalized avatars for different identities. Following GaussianAvatars [39], we train on all 15 cameras except the frontal and report results on the frontal camera for all our experiments. All images are resized to 1604×1100 during training. For avatar reconstruction, we use 30 short sequences. For audio-driven animation, we use 300 sequences for training and 50 for val and test set each. We encourage readers to watch the Supplementary Video for visual comparison of all results.

5.1. Avatar Reconstruction

Compared to GaussianAvatars [39], our proposed avatar initialization can generate high-quality results with as few as 30-35k points (see Fig. 5 and Tab. 2). The perceptual loss helps increase the sharpness in the texture with fewer points. The wrinkle regularization helps to model dynamic wrinkles. Teeth subdivision helps with the better mouth interior. Color MLP helps synthesize sharper texture. Our full avatar initialization with all regularization achieves the best results. We train our method on all except frontal camera and report results for the frontal camera. For these experiments, we show results for the most expressive actor from our dataset (Subject 4) and refer to Suppl. doc for others.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	# Gaussians ↓
GaussianAvatar [39]	26.53	0.9087	0.1487	98083
Ours (w/o perceptual)	27.03	0.9116	0.1447	**31875**
Ours (w/o wrinkle reg.)	28.10	0.9216	0.1312	33998
Ours (w/o mouth subdivision)	28.35	0.9321	0.1244	34917
Ours (w/o Color MLP)	28.93	0.9366	0.1235	32792
Ours (Full)	**29.90**	**0.9495**	**0.1104**	32379

Table 2. Avatar Reconstruction: With fewer Gaussian points, our method achieves superior quality compared to the alternate approaches. Perceptual loss increases the sharpness, wrinkle regularization models dynamic wrinkles, mouth subdivision learns better mouth interior, Color MLP synthesizes sharper colors and accurate dynamic wrinkles. The full avatar initialization with all regularizations achieves the best results.

	Method	LSE-D ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NeRF	RAD-NeRF [49]	13.17	13.15	0.8007	0.2741
	ER-NeRF [29]	13.08	15.94	0.8269	0.2512
	SyncTalk [37]	12.50	18.24	0.8759	0.1920
3DGS	TalkingGaussian [30]	12.38	20.29	0.8890	0.1745
	GaussianTalker [8]	12.19	20.32	0.8984	0.1724
FLAME	Faceformer [15] + G.A.	11.86	22.18	0.9105	0.1608
	CodeTalker [63] + G.A.	11.68	22.23	0.9118	0.1595
	Imitator [51] + G.A.	11.61	22.83	0.9207	0.1519
	Ours	**11.25**	**24.73**	**0.9362**	**0.1286**

Table 3. Baseline Comparisons: we compare with NeRF-based, 3DGS-based and mesh-based (FLAME [31]) baselines. We combine FLAME-based methods with 3DGS via GaussianAvatars (G.A.) [39]. Our method achieves superior results in both in perceptual quality as well as lip synchronization (LSE-D).

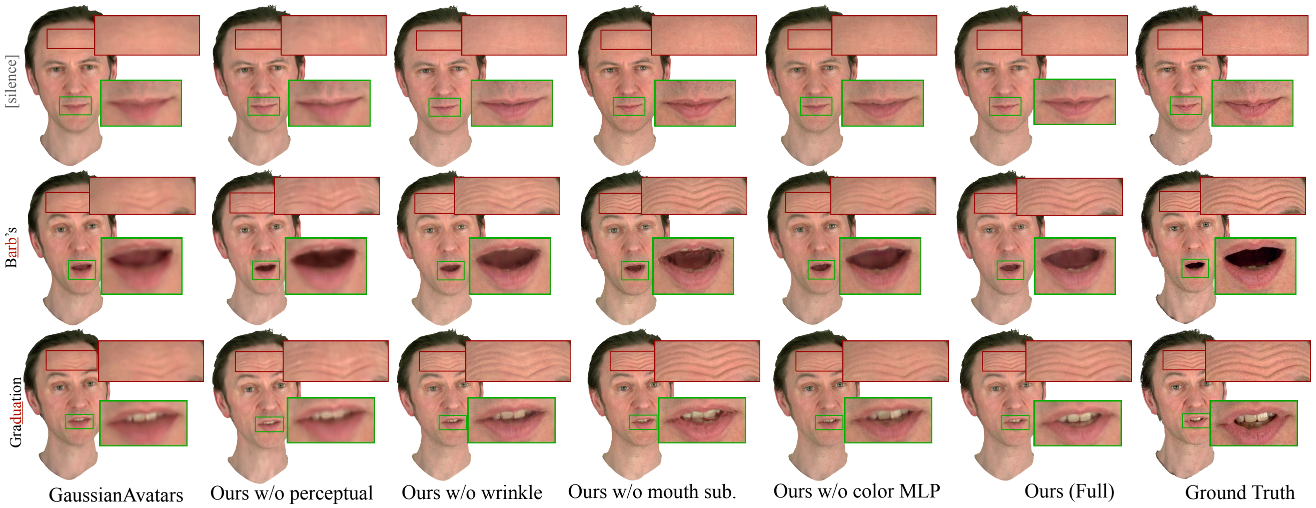

Figure 5. Avatar Reconstruction: GaussianAvatars [39] produces blurry results and cannot handle dynamic wrinkles. For our method, without perceptual loss it cannot synthesize sharp textures for global & local less observed regions like teeth, wrinkle regularization helps to model dynamic wrinkles, mouth faces subdivision helps with the better mouth interior and Color MLP helps synthesize sharper colors and accurate dynamic wrinkles. Our full avatar initialization technique with all regularization achieves the best results.

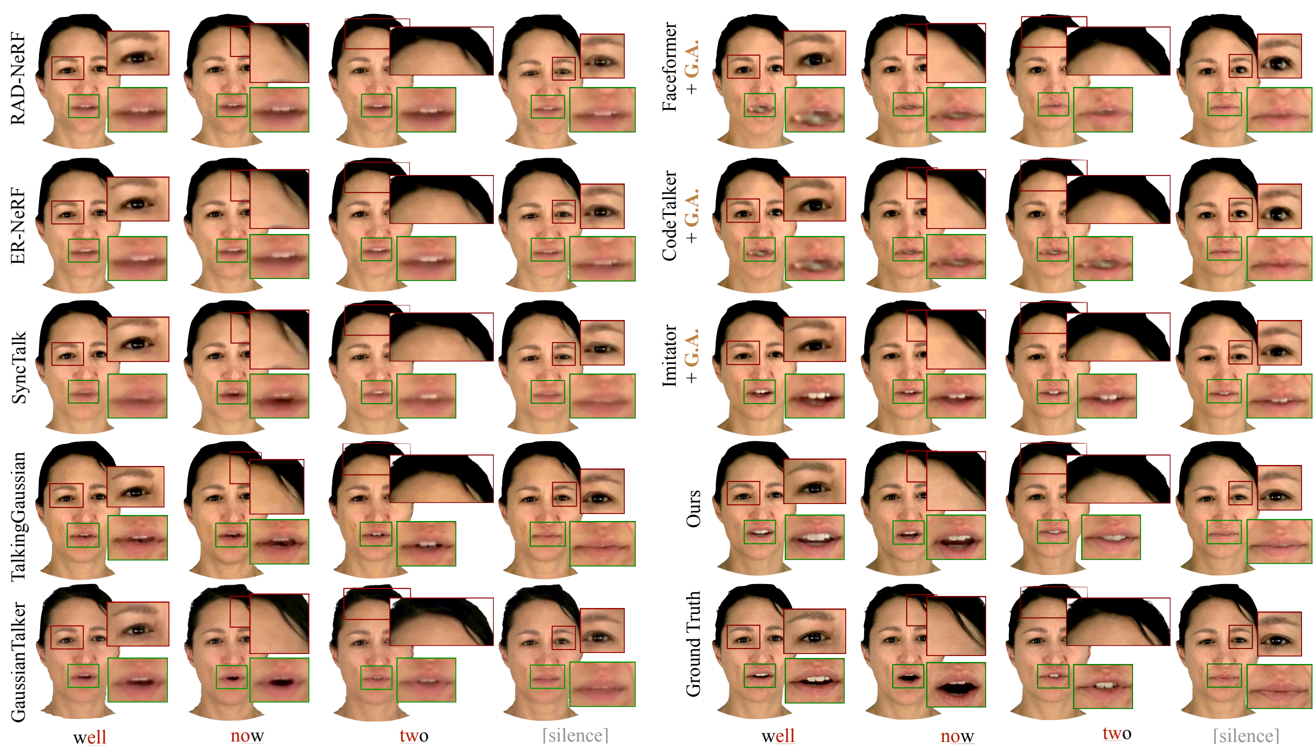

Figure 6. Baseline Comparison: We show comparisons against NeRF-based, 3DGS-based and FLAME animation methods combined with GaussianAvatars (G.A.) [39]. NeRF-based methods (RAD-NeRF [49], ER-NeRF [29] and SyncTalk [37]) produce artifacts in texture as well as incorrect mouth articulations. 3DGS-based methods (TalkingGaussian [30] & GaussianTalker [8]) can synthesize better lip-sync but produces blurry texture especially for mouth interior. Generalized FLAME animation methods (Faceformer [15], CodeTalker [63]) show blurred mouth interiors, personalized methods (Imitator [51]) produce better mouth interiors, however, the lip closures and synchronization is inaccurate. Our method outperforms all baselines both in lip-sync and photorealism.

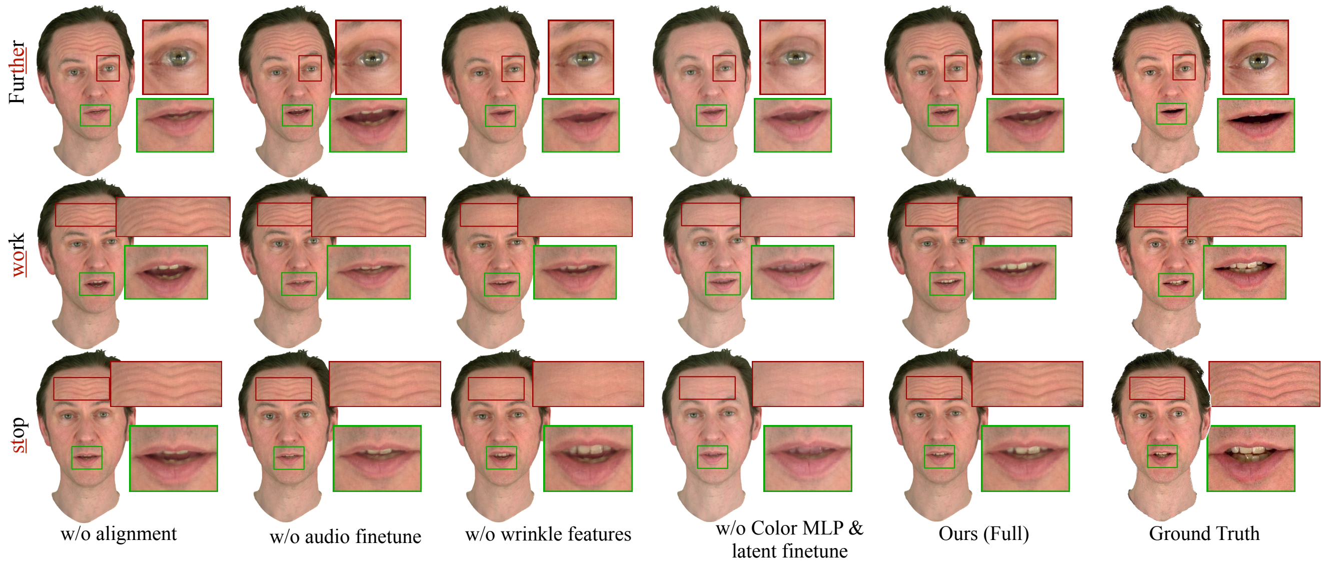

Figure 7. Ablation Study. Left-to-right: (1) Alignment mask is critical to properly infuse audio information into the sequence model. (2) Audio fine-tuning helps the method generate better lip sync. (3) Without wrinkle features, the model can not produce dynamic wrinkles. (4) Without fine-tuning Color MLP and latent features, the model produces bad mouth interiors and inaccurate dynamic wrinkles. Our full model with all the components achieves best results.

5.2. Audio-Driven Animation

Baseline Comparisons. We compare our method against recent state-of-the-art methods. For NeRF- and 3DGS-based methods, we train on frontal camera since these methods are designed for monocular settings. There are no sequence models for audio-driven animation of 3D head avatars, thus, we combine audio-to-mesh animation methods [15, 51, 63] with current state-of-the-art mesh-to-3D avatar creation method [39]. We report results on the front camera for fairness, since some methods are designed only for front/single camera only. We report results averaged over all subjects, see Fig. 6 and Tab. 3. Our method consistently achieves better results than baselines both in terms of perceptual quality and lip synchronization.

Method	LSE-D ↓	PSNR ↑	SSIM ↑	LPIPS ↓
w/o alignment	12.66	21.02	0.9104	0.1855
w/o audio finetune	11.78	22.73	0.9355	0.1198
w/o wrinkle features	11.28	23.14	0.9311	0.1162
w/o color MLP & latent finetune	11.32	23.96	0.9367	0.1133
Ours (Full)	**11.15**	**24.97**	**0.9470**	**0.1101**

Table 4. Ablation study. Without alignment mask, the model ignores the audio signal. Audio fine-tuning helps to improve lip sync. Wrinkle features help with dynamic wrinkles and overall realism. Finetuning Color MLP and latents rectifies the inaccurate mouth interior. Our full model achieves the best results.

Ablation Study. Finally, we ablate different design choices of our method on most expressive actor from our dataset (Subject 4) in Fig. 7 and Tab. 4. Alignment mask is critical for accurately infusing audio features into the sequence model. Without audio fine-tuning refers to using generic audio features without any personalization of lip encoder, without audio model fine-tuning the model produces incorrect lip synchronization. Without wrinkle features refers to setting without using wrinkle features for producing FLAME expressions. Without wrinkle features the method cannot produce dynamic wrinkles. Without fine-tuning Color MLP & latent features with predicted expressions from our Expression encoder, the method produces bad mouth interiors and inaccurate dynamic wrinkles. Our full model with all components achieves best results. We refer readers to supplemental video for visual comparison.

6. Conclusion

In this work, we propose a novel approach to create high-fidelity and photorealistic 3D head avatars that can be animated from audio input. We designed the first transformer-based sequence model for audio-driven head animation of 3DGS based avatar. Our sequence model is made possible by a lightweight and compact avatar initialization based on 3D Gaussian Splatting. We proposed several regularization techniques to handle dynamic wrinkles, skin creasing and sharpness of the texture. Our method produces (a) photorealistic and high-quality 3D head avatars that can be rendered from arbitrary viewpoints (b) visually natural animations like skin creasing during talking. We believe this is an important first step towards enabling the animation of detailed and lightweight 3D head avatar, which can enable many new possibilities for content creation and digital avatars for immersive telepresence.

7. Acknowledgments

This work was supported by the ERC Starting Grant Scan2CAD (804724), the Bavarian State Ministry of Science and the Arts and coordinated by the Bavarian Research Institute for Digital Transformation (bidt), the German Research Foundation (DFG) Grant "Making Machine Learning on Static and Dynamic 3D Data Practical," the German Research Foundation (DFG) Research Unit "Learning and Simulation in Visual Computing". We would like to thank Shenhan Qian for help with tracking.

References

[1] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. Facetalk: Audio-driven motion diffusion for neural parametric head models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[2] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. 4, 5

[3] Yunpeng Bai, Yanbo Fan, Xuan Wang, Yong Zhang, Jingxiang Sun, Chun Yuan, and Ying Shan. High-fidelity facial avatar reconstruction from monocular video with generative priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4541–4551, 2023. 1

[4] Lele Chen, Zhiheng Li, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance, 2018. 2

[5] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7832–7841, 2019.

[6] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. *arXiv preprint arXiv:2007.08547*, 2020.

[7] 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*, New York, NY, USA, 2022. Association for Computing Machinery. 2

[8] Kyusun Cho, Joungbin Lee, Heeji Yoon, Yeobin Hong, Jaehoon Ko, Sangjun Ahn, and Seungryong Kim. Gaussiantalker: Real-time talking head synthesis with 3d gaussian splatting. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 10985–10994, New York, NY, USA, 2024. Association for Computing Machinery. 1, 2, 6, 7

[9] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that?, 2017. 2

[10] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael Black. Capture, learning, and synthesis of 3D speaking styles. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10101–10111, 2019. 2

[11] Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. Emotional speech-driven animation with content-emotion disentanglement. In *SIGGRAPH Asia 2023 Conference Papers*, New York, NY, USA, 2023. Association for Computing Machinery. 2

[12] Dipanjan Das, Sandika Biswas, Sanjana Sinha, and Brojeshwar Bhowmick. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, page 408–424, Berlin, Heidelberg, 2020. Springer-Verlag. 2

[13] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*, page 741–757, Berlin, Heidelberg, 2020. Springer-Verlag. 4

[14] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14398–14407, 2021. 2

[15] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 6, 7, 8

[16] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. Darpa timit acoustic phonetic continuous speech corpus cdrom, 1993. 3

[17] Jiazhi Guan, Zhanwang Zhang, Hang Zhou, Tianshu HU, Kaisiyuan Wang, Dongliang He, Haocheng Feng, Jingtuo Liu, Errui Ding, Ziwei Liu, and Jingdong Wang. Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[18] Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[19] Anchit Gupta, Rudrabha Mukhopadhyay, Sindhu Balachandra, Faizan Farooq Khan, Vinay P. Namboodiri, and C. V. Jawahar. Towards generating ultra-high resolution talking-face videos with lip synchronization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5209–5218, 2023. 2

[20] Siddharth Gururani, Arun Mallya, Ting-Chun Wang, Rafael Valle, and Ming-Yu Liu. Space: Speech-driven portrait animation with controllable expression, 2022. 2

[21] Qianyun He, Xinya Ji, Yicheng Gong, Yuanxun Lu, Zhengyu Diao, Linjia Huang, Yao Yao, Siyu Zhu, Zhan Ma, Songchen Xu, Xiaofei Wu, Zixiao Zhang, Xun Cao, and Hao Zhu. Emotalk3d: High-fidelity free-view synthesis of emotional 3d talking head. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 3

[22] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

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

[24] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016. 4

[25] Prajwal K R, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and C V Jawahar. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, page 1428–1436, New York, NY, USA, 2019. Association for Computing Machinery. 2

[26] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graph.*, 36 (4), 2017. 2

[27] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 2, 5, 6

[28] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 42(4), 2023. 2, 3

[29] 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 (ICCV)*, pages 7568–7578, 2023. 1, 2, 6, 7

[30] 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 *Computer Vision – ECCV 2024*, pages 127–145, Cham, 2025. Springer Nature Switzerland. 1, 2, 6, 7

[31] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 4, 6

[32] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-aware implicit neural audio-driven video portrait generation. *arXiv preprint arXiv:2201.07786*, 2022. 2

[33] Steven R. Livingstone and Frank A. Russo. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), 2018. 3

[34] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis, 2020. 2

[35] Soumik Mukhopadhyay, Saksham Suri, Ravi Teja Gadde, and Abhinav Shrivastava. Diff2lip: Audio conditioned diffusion models for lip-synchronization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5292–5302, 2024. 2

[36] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20687–20697, 2023. 2

[37] Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, 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 (CVPR)*, 2024. 1, 2, 6, 7

[38] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. 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*, page 484–492, New York, NY, USA, 2020. Association for Computing Machinery. 2, 6

[39] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. *arXiv preprint arXiv:2312.02069*, 2023. 3, 4, 6, 7, 8

[40] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando de la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1173–1182, 2021. 1, 2

[41] Shrimanta Satpati. Wrinkle Detection Streamlit. `https://github.com/shrimantasatpati/Wrinkle-Detection-StreamLit`, 2023. 4, 5

[42] 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 *ECCV*, 2022. 1, 2

[43] Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *CVPR*, 2023. 2

[44] Shuai Shen, Wanhua Li, Xiaoke Huang, Zheng Zhu, Jie Zhou, and Jiwen Lu. Sd-nerf: Towards lifelike talking head animation via spatially-adaptive dual-driven nerfs. *IEEE Transactions on Multimedia*, 26:3221–3234, 2024. 1

[45] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want, 2020. 2

[46] Michał Stypułkowski, Konstantinos Vougioukas, Sen He, Maciej Zikeba, Stavros Petridis, and Maja Pantic. Diffused Heads: Diffusion Models Beat GANs on Talking-Face Generation. In *https://arxiv.org/abs/2301.03396*, 2023. 2

[47] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph.*, 2017. 2

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

[49] 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. 6, 7

[50] Balamurugan Thambiraja, Sadegh Aliakbarian, Darren Cosker, and Justus Thies. 3diface: Diffusion-based speech-driven 3d facial animation and editing, 2023. 2

[51] Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20621–20631, 2023. 1, 2, 6, 7, 8

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

[53] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive - generating expressive portrait videos with audio2video diffusion model under weak conditions, 2024. 2

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 5

[55] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven facial animation with temporal gans, 2018. 2

[56] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans, 2019.

[57] Jiayu Wang, Kang Zhao, Shiwei Zhang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Lipformer: High-fidelity and generalizable talking face generation with a pre-learned facial codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13844–13853, 2023. 2

[58] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020. 2, 3

[59] Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. One-shot talking face generation from single-speaker audio-visual correlation learning. In *AAAI 2022*, 2022. 2

[60] O. Wiles, A.S. Koepke, and A. Zisserman. X2face: A network for controlling face generation by using images, audio, and pose codes. In *European Conference on Computer Vision*, 2018. 2

[61] Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Xuhua Huang, Alexander Hypes, Taylor Koska, Steven Krenn, Stephen Lombardi, Xiaomin Luo, Kevyn McPhail, Laura Millerschoen, Michal Perdoch, Mark Pitts, Alexander Richard, Jason Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shoou-I Yu, and Yaser Sheikh. Multiface: A dataset for neural face rendering. In *arXiv*, 2022. 3

[62] You Xie, Hongyi Xu, Guoxian Song, Chao Wang, Yichun Shi, and Linjie Luo. X-portrait: Expressive portrait animation with hierarchical motion attention. 2024. 2

[63] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12780–12790, 2023. 1, 2, 6, 7, 8

[64] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time, 2024. 2

[65] Shunyu Yao, RuiZhe Zhong, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering. *arXiv preprint arXiv:2201.00791*, 2022. 1, 2

[66] Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, Jinzheng He, and Zhou Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. *arXiv preprint arXiv:2301.13430*, 2023. 1, 2

[67] 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. 2

[68] Zicheng Zhang, Ruobing Zheng, Ziwen Liu, Congying Han, Tianqi Li, Meng Wang, Tiande Guo, Jingdong Chen, Bonan Li, and Ming Yang. Learning dynamic tetrahedra for high-quality talking head synthesis, 2024. 1

[69] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 2

[70] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: Speaker-aware talking-head animation. *ACM Trans. Graph.*, 39(6), 2020. 2