GaussianSpeech: Audio-Driven Personalized 3D Gaussian Avatars

Supplementary Material

In this supplemental document, we provide additional results of our proposed method GaussianSpeech, including a user study, in Sec. A. Details about our network architecture and training scheme are given in Sec. B, including preliminaries used in the main paper, see Sec. C. In Sec. D, we have added a further discussion of the baseline methods, and in Sec. E we provide more dataset details.

A. Additional Experiments

A.1. Novel View Synthesis

In our experiments, we found that training with at least 30 sequences enables generalizing for mouth articulations. We show novel view synthesis results and zoom-ins in Fig. 1. For all the avatars from our dataset, we show results for avatar initialization in Fig. 2 and Tab. 1. Audio-driven animations are shown in Fig. 3.

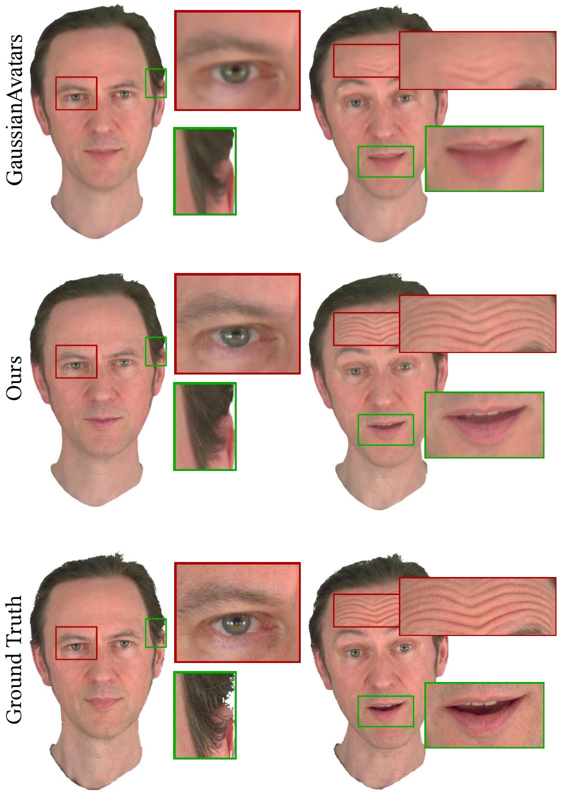


Figure 1. Novel View Synthesis quality in comparison to GaussianAvatars [16]. In contrast to ours, GaussianAvatars generates blurry texture and cannot generate dynamic wrinkles.

Subjects	PSNR ↑	SSIM ↑	LPIPS ↓	# Gaussians ↓
Subject 1	30.07	0.9598	0.0754	30653
Subject 2	30.13	0.9180	0.1125	32443
Subject 3	28.62	0.9607	0.1072	31434
Subject 4	29.90	0.9495	0.1104	32379
Subject 5	30.05	0.9529	0.1185	29490
Subject 6	26.82	0.9228	0.1322	35138

Table 1. Avatar Initialization: All avatars converge in the range of 29-35K Gaussians. We show perceptual quality metrics for each of our avatars evaluated for novel views.

A.2. Effect of Latent Features

We analyze the effect of per Gaussian latent features during our avatar initialization stage in Fig. 4. Note that the per Gaussian features are critical to produce accurate texture colors for the avatar.

A.3. Failure Cases

While our method produces photorealistic and high-quality animations in synchronization with audio, it also has several limitations. Our avatar initialization strategy is based on FLAME [13]; thus, our method struggles with avatars wearing accessories like glasses. The glass geometry and specularities on the surface of glasses can not be accurately produced and fails during free-viewpoint rendering, see Fig. 5. In the future, this can be improved by designing better models for representing human head geometry instead of 3D mesh. Also, our texture representation based on the Color MLP has baked-in lighting, and cannot be separated from material properties, which is important for placing avatars in different environments (e.g., during immersive telepresence).

A.4. User Study

To evaluate the fidelity based on human perceptual evaluation, we performed a user study with 30 participants over a set of 15 questions. The users were given a carefully crafted set of instructions to evaluate (a) Overall Animation Quality (b) Lip Synchronization and (c) Realism in Facial Movements. The users were asked to assess different anonymous methods (including GaussianSpeech) on these three parameters. In the course of the study, participants were presented with these questions to focus on different aspects of 3D facial animation, shown in Fig 6. For every question, participants were instructed to meticulously evaluate the provided methods and select the option that best aligned with their judgment.

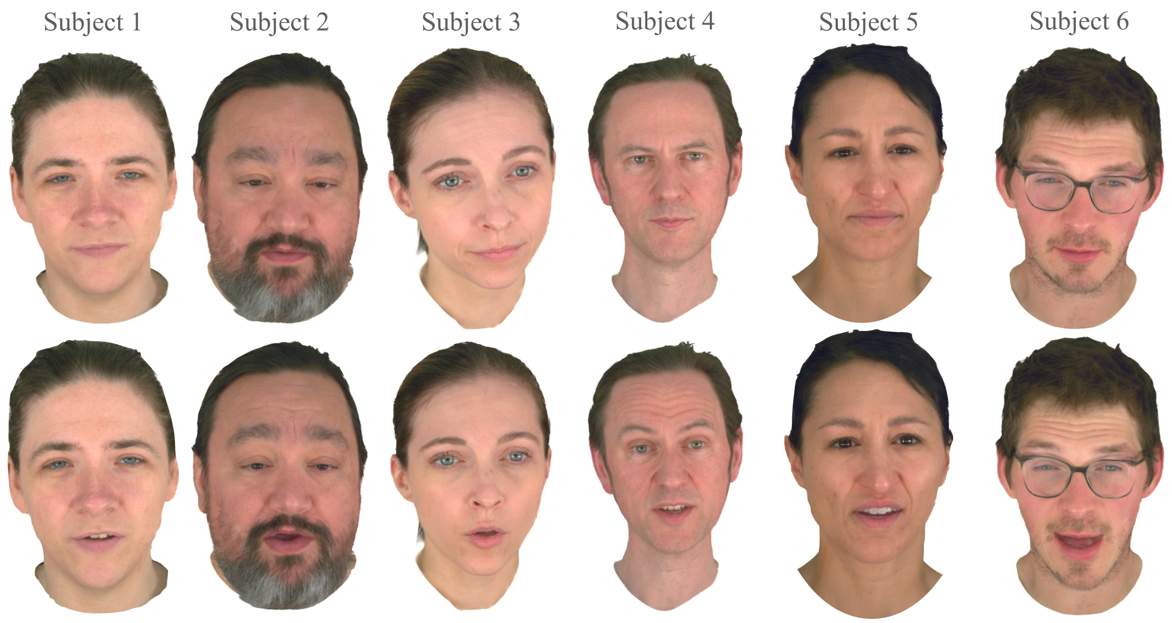


Figure 2. Reconstructed Avatars: We show the reconstructed avatars from novel views for all the participants from our dataset. We visualize two randomly selected frames for each avatar, one where the avatar is silent and the other where the avatar is speaking. Our method generates high quality avatars generating realistic and sharper textures.

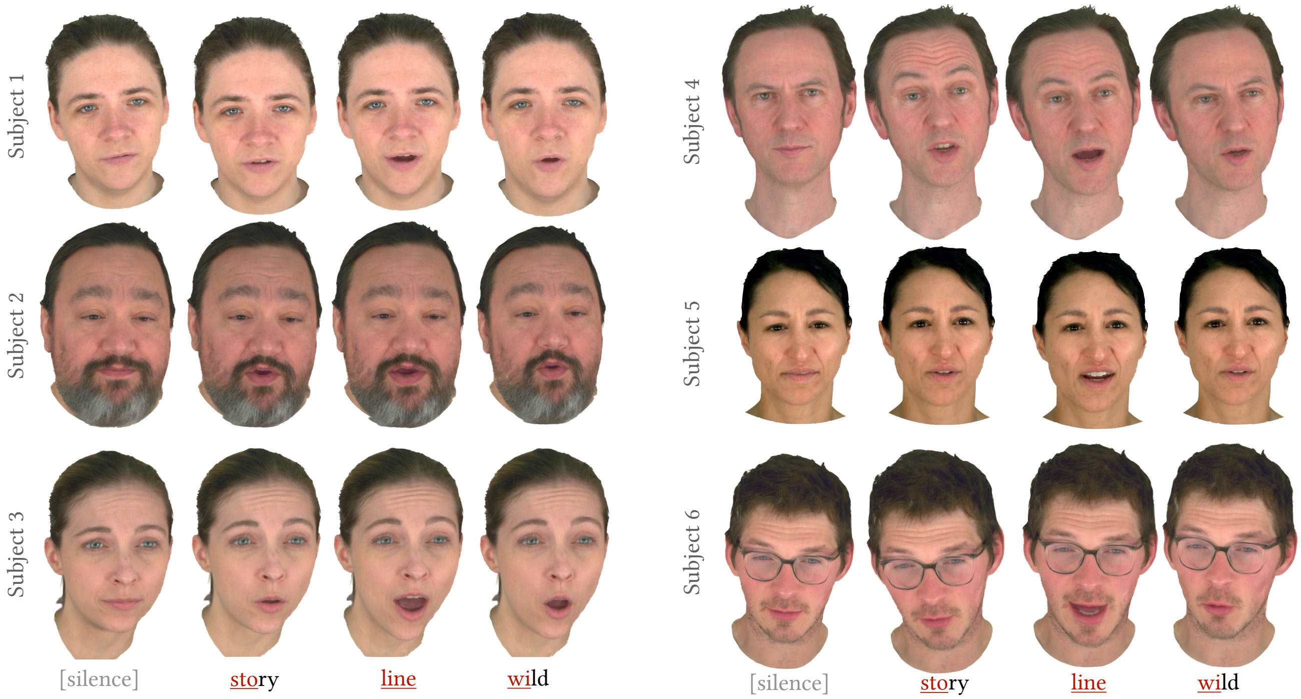


Figure 3. Audio-Driven Animation: We show animation results of our avatars animated directly from audio signal for the novel camera. The words spoken by the avatars are highlighted at the bottom.

For the first question evaluating overall quality, participants were instructed to consider factors such as visual appeal, clarity, and general impression, and to choose the method number that they believed demonstrates the highest overall quality. For the second question, participants were directed to evaluate the lip synchronization of each anima-

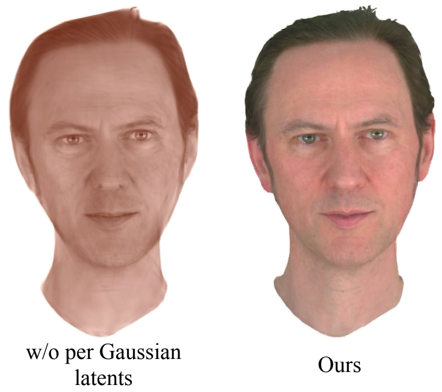


Figure 4. Effect of per Gaussian latents: Without the features, the color MLP cannot produce accurate texture colors.

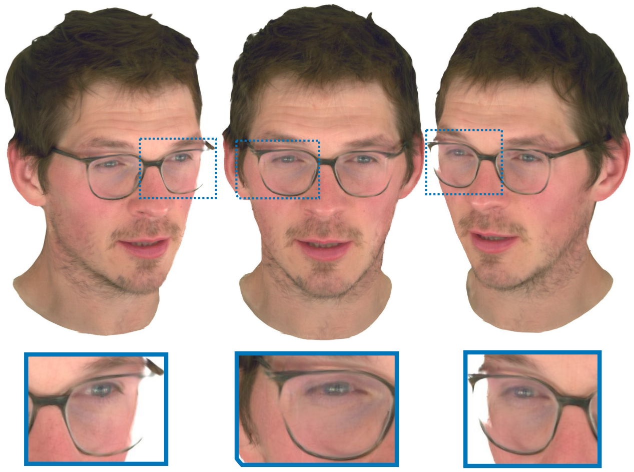
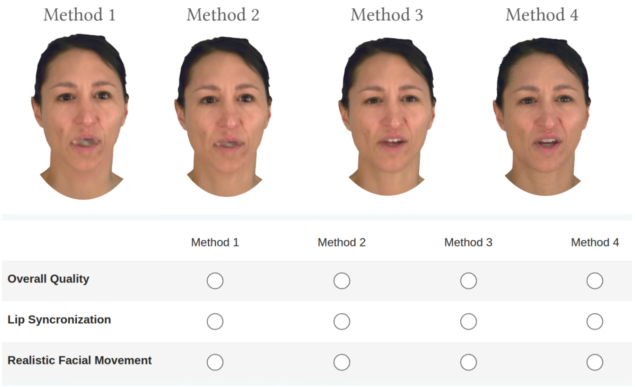


Figure 5. Failure Cases: The method fails to produce accurate accurate glass geometry and specularities on the surface of glasses.

tion method. They were prompted to pay close attention to how well the lip movements aligned with the spoken words or sounds. Participants were reminded to select only one option that, in their judgment, exhibited the best lip synchronization. Lastly, the third question was focussed on evaluating the realistic facial movement of each 3D facial animation method. Participants were instructed to consider the naturalness and persuasiveness of facial expressions and movements and to choose the method number that, in their opinion, demonstrates the most realistic facial movement. Again, participants were reminded to select only one option per question throughout the study.

Our method consistently achieves better lip-audio synchronization while also representing fine-scale facial details like skin creasing and wider mouth motions. This is confirmed by our perceptual user study in Fig. 7.



	Method 1	Method 2	Method 3	Method 4
Overall Quality	○	○	○	○
Lip Syncronization	○	○	○	○
Realistic Facial Movement	○	○	○	○

Figure 6. Different methods shown to users and questions asked to assess the quality of different methods during perceptual study evaluation. Method names were anonymized to avoid bias towards a particular method. Users were asked to select one of the shown methods for each of the three questions based on which one they believe exhibits best results.

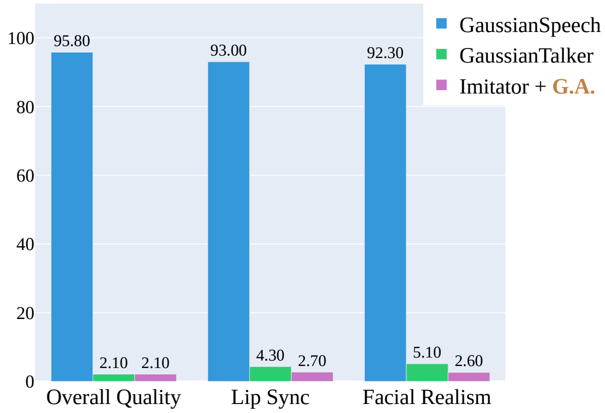


Figure 7. User study comparison with baselines. We measure preference for (1) Overall Animation Quality, (2) Lip Synchronization and (3) Facial Realism. GaussianSpeech results are overwhelmingly preferred over the best baseline methods on all these aspects.

A.5. Inference Speed

We report inference speed averaged over the test set on a single Nvidia RTX 2080 Ti with 12GB VRAM as well as NVIDIA RTX A6000 with 48GB VRAM. Since TalkingGaussian [12] network does not fit in 12 GB VRAM; we report its inference time only for a 48 GB VRAM (NVIDIA A6000). The results are presented in Tab. 2.

A.6. Results on In-the-Wild Monocular Videos

We first run monocular face tracking to obtain tracked FLAME meshes. Next, we use HuBERT encoder (similar to baselines) to extract audio features and use it to train our transformer decoder model. Although the main focus of GaussianSpeech is audio-driven 3D avatars, it can still pro-

Method	FPS (2080Ti)↑	FPS (A6000)↑	# Gaussians
Faceformer [7] + G.A.	25.38	42.23	65-100K
CodeTalker [21] + G.A.	23.92	38.98	65-100K
Imitator [20] + G.A.	23.32	39.71	65-100K
SyncTalk [14]	10.14	21.51	N/A
ER-NeRF [11]	16.88	17.98	N/A
RAD-NeRF [19]	17.18	21.41	N/A
GaussianTalker [3]	57.82	59.01	41K-44K
TalkingGaussian [12]	OOM	73.34	31K-98K
Ours	**74.29**	**123.48**	30-35K

Table 2. Inference Speed on Nvidia RTX 2080 Ti (12GB VRAM) and NVIDIA A6000 (48GB VRAM). TalkingGaussian's network does not fit in 12 GB VRAM and throws Out-of-Memory (OOM) error; for this method, the inference time is only given for a 48 GB VRAM (NVIDIA A6000).

duce high-quality results for monocular videos; see comparisons in Fig. 8 and Tab. 3.

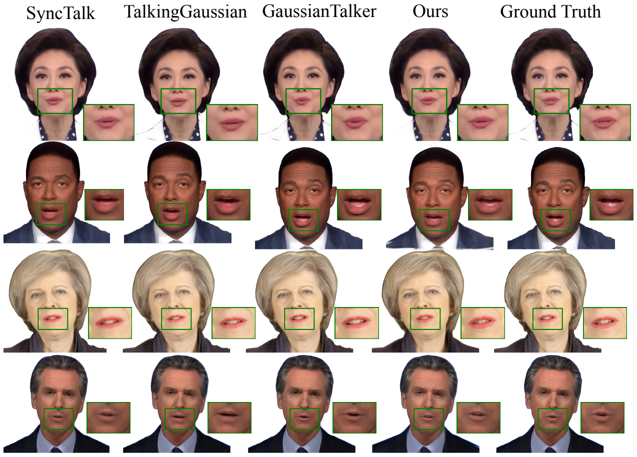


Figure 8. Comparison against monocular in-the-wild baselines.

Method	LSE-D↓	PSNR↑	SSIM↑	LPIPS↓
SyncTalk [14]	11.83	29.34	0.9136	0.0550
TalkingGaussian [12]	11.68	30.53	0.9477	0.0336
GaussianTalker [3]	11.62	31.01	**0.9481**	0.0316
Ours	**11.56**	**31.53**	0.9366	**0.0289**

Table 3. GaussianSpeech outperforms SyncTalk and TalkingGaussian, with perceptual improvement over GaussianTalker..

B. Architecture & Training Details

GaussianSpeech is implemented using PyTorch Lightning framework [6] with Wandb [2] for logging.

B.1. Implementation Details

Avatar Representation. During avatar initialization, the teeth subdivision changes the initial mesh with 5143 vertices and 10144 faces to 5431 vertices and 10648 faces. For

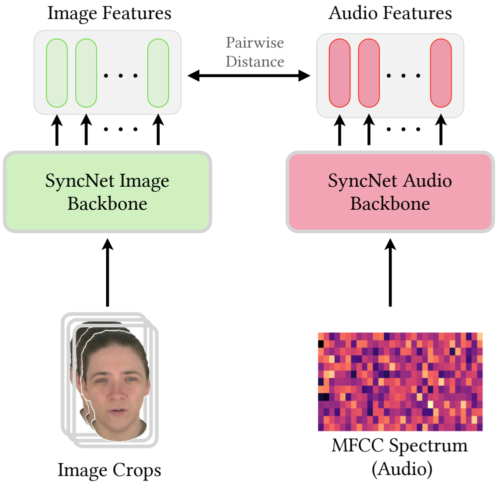


Figure 9. The image crops are passed to the pretrained Syncnet image backbone to extract image features, and audio features, represented as MFCC power spectrum, are extracted via pretrained Syncnet audio backbone. Finally, the pairwise distance between image and audio features are calculated to compute lip synchronization.

perceptual loss $\mathcal{L}_{\text{global}}$, we downscale images to 401×275. The volume based pruning ensures that the avatar converges to 30-35k Gaussian points. We use Adam optimizer with exponential decay and the default learning rates from [16]. For the loss (Eq. **??** main paper), we use $\lambda_{\text{pos}} = 0.01$, $\lambda_{\text{s}} = 1$, $\lambda_{\text{g}} = 1$, $\lambda_{\text{p}} = 0.001$ and $\lambda_{\text{w}} = 10$. For rendering, we use the differentiable tile rasterizer [9].

Audio-Driven Sequence Generation. For the first 10K iterations, we train the Lip transformer, Wrinkle transformer and Expression encoder each, and then use them in our training pipeline. For the vertex loss $\mathcal{L}_{\text{vertices}}$ (Eq. **??** main paper), we predict 5431 offsets obtained after mouth subdivision. For the next 2000 iterations, we train the model only with $\mathcal{L}_{\text{vertices}}$ to ensure that it learns coarse motion reasonably in coherence with the audio. And then gradually add $\mathcal{L}_{\text{photo}}$, and train the model together with $\mathcal{L}_{\text{vertices}}$ and $\mathcal{L}_{\text{photo}}$ in an alternating fashion.

B.2. Architecture Details

Avatar Initialization. During avatar initialization (Sec. **??**, main paper), we randomly sample 16 random patches per iteration with a patch size of 128×128 from the facial area. We start with an initial learning rate of 5e-3 and exponentially decay until 5e-5. We perform densification every 5000 iterations and we do not reset opacity. We train with the batch size of 1 with Adam optimizer, render images on white background and train for 100,000 iterations on RTX 2080 Ti (12 GB VRAM). Our avatars converge between 29-35K Gaussian points, as shown in Tab. 1. We show sample output of our wrinkle detection model used in main paper in Fig. 10.

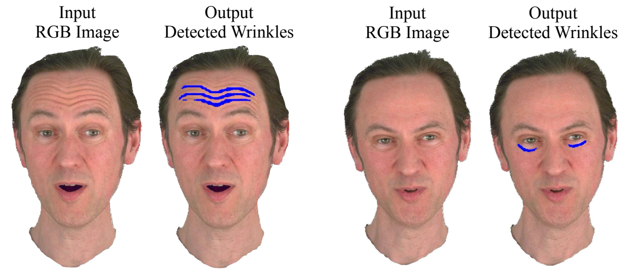


Figure 10. Sample output of the wrinkle detection model [18] used for wrinkle regularization loss during avatar training.

Audio-to-Avatar Model. For the audio encoder, the TCN layers of the Wav2Vec 2.0 [1] are initialized with the pre-trained wav2vec 2.0 weights trained on a large corpus of audio data from different languages and is frozen during fine-tuning. The Frequency Interpolation layer simply performs linear interpolation of the incoming features and has no learnable parameters. For the Lip an Wrinkle encoder, we use latent dimension of 64 and for Expression encoder and the Transformer decoder the latent dimension is 128. For the multihead self and cross attention layers of the transformer decoder, we use 4 heads and set the dimension to 1024 for each decoder block. We use an Adam optimizer with a learning rate of 1e-4 and update the model one sequence per iteration, train on Nvidia RTX A6000 (48 GB VRAM) for 100,000 iterations.

Evaluation Metrics. To evaluate lip synchronization of the generated mouth expressions with the audio signal, we use LSE-D (Lip Sync Error Distance) [15]. Specifically, this involves feeding rendered face crops and the corresponding audio signal into a pre-trained SyncNet [4] to evaluate how close the acoustic signal matches the phonetic movements. The facial movements are encoded as crops of only the facial region, and the audio signal is represented as MFCC power spectrum. These are then passed into the pretrained SyncNet backbone [4] and the pairwise distance is evaluated, as shown in Fig. 9.

C. Preliminaries

C.1. 3D Gaussian Splatting

Recently, 3D Gaussian Splatting [9] has emerged as a promising approach to represent a static scene explicitly with anisotropic 3D gaussian directly from multiview images and estimated/given camera poses. Specifically, it represents a scene using a set of 3D Gaussian splats, each defined by a set of optimizable parameters, including a mean position $\boldsymbol{\mu} \in \mathbb{R}^3$ and a positive semi-definite covariance matrix $\Sigma \in \mathbb{R}^{3\times 3}$ as:

$$G(\boldsymbol{x}) = e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \tag{1}$$

Given that covariance matrix Σ needs to be positive semidefinite to have physical meaning and gradient-based optimization methods cannot be constrained to produce such valid matrices, Kerbl *et al.* [9] first define an ellipsoid with scaling matrix S and rotation matrix R as:

$$\Sigma = RSS^TR^T. \tag{2}$$

To allow for independent optimization for scale and rotation, separate 3D vectors are stored. The scale is represented using a scaling vector $\boldsymbol{s} \in \mathbb{R}^3$ and a quaternion $\boldsymbol{q} \in \mathbb{R}^4$ for rotation.

For rendering every pixel on the image, the color $\boldsymbol{C}$ is computed by blending all the 3D Gaussians overlapping a pixels as:

$$\boldsymbol{C} = \sum_{i=1}^{N} \boldsymbol{c}_i\alpha_i \prod_{j=1}^{i-1}(1-\alpha_j) \tag{3}$$

where $\boldsymbol{c}_i$ refers to 3-degree spherical harmonics (SH) [17] color obtained by blending N ordered points overlapping the pixel, blending weight α_i is given by multiplying 2D projection of the 3D Gaussian with learnt per-point opacity. Paired with a differentiable tile rasterizer, this enables real-time rendering. To handle complex scenes and respect visibility order, depth-based sorting is applied to the Gaussian splats before blending.

C.2. GaussianAvatars

Due to the capability of 3DGS to represent fine geometric structures, it has proved to be an efficient representation for creating photorealistic head avatars, as shown by GaussianAvatars [16]. GaussianAvatars proposes a method for dynamic 3D representation of human heads based on 3DGS by rigging the anisotropic 3D Gaussians to the faces of a 3D morphable face model. Specifically, the method uses FLAME [13] as 3DMM due to its flexibility and compactness, consisting of only 5023 vertices and 9976 faces. To better represent mouth interior, it generated additional 120 vertices for teeth. Given a FLAME mesh, the idea is to first initialize a 3D Gaussian at the center of each triangle of the FLAME mesh and let the 3D Gaussian move with the faces of the FLAME mesh across different timesteps. For the paired 3D Gaussians with the faces of the FLAME mesh, the position $\boldsymbol{\mu}$, rotation $\boldsymbol{r}$ and anisotropic scaling $\boldsymbol{s}$ are defined in local space. During rendering, these are converted to global space as:

$$\boldsymbol{r}' = \boldsymbol{R}\boldsymbol{r}, \tag{4}$$

$$\boldsymbol{\mu}' = k\boldsymbol{R}\boldsymbol{\mu} + \boldsymbol{T} \tag{5}$$

$$\boldsymbol{s}' = k\boldsymbol{s}, \tag{6}$$

where $\boldsymbol{T}$ refers to the mean positions of the vertices of the triangle mesh, rotation matrix $\boldsymbol{R}$ describes the orientation of the triangles in the global space, scalar k describes the triangle scaling. During avatar optimization, similar to 3DGS, the method uses adaptive density control strategy to add and remove splats based view-space positional gradient and opacity of each Gaussian. To prevent excessive pruning, the method also ensures that every triangle has at least one splat attached. The 3DGS parameters are then optimized using photometric loss $\mathcal{L}_{\text{rgb}}$, position loss $\mathcal{L}_{\text{position}}$ and scaling loss as $\mathcal{L}_{\text{scale}}$ as:

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{pos}}\mathcal{L}_{\text{position}} + \lambda_{\text{s}}\mathcal{L}_{\text{scaling}} \tag{7}$$

where $\mathcal{L}_{\text{rgb}}$ is combination of $\mathcal{L}_1$ and D-SSIM loss [9] between rendered and ground truth images.

$$\mathcal{L}_{\text{rgb}} = (1-\lambda)\mathcal{L}_1 + \lambda\mathcal{L}_{\text{D-SSIM}}, \tag{8}$$

$\mathcal{L}_{\text{position}}$ ensures that splats remain close to their parent triangles:

$$\mathcal{L}_{\text{position}} = \big|\big|\max(\boldsymbol{\mu}, \boldsymbol{\epsilon}_{\text{position}})\big|\big|_2, \tag{9}$$

and $\mathcal{L}_{\text{scaling}}$ prevents excessive scaling of the splats:

$$\mathcal{L}_{\text{scaling}} = \big|\big|\max(\boldsymbol{s}, \boldsymbol{\epsilon}_{\text{scaling}})\big|\big|_2, \tag{10}$$

D. Baselines

We compare our method against audio-conditioned NeRF, 3DGS and mesh based methods. Since mesh-based methods can't generate photorealistic avatars, we combined audio-to-mesh methods with recent state-of-the-art mesh-to-3D avatar method GaussianAvatars [16]. Current NeRF and 3DGS based methods are designed for monocular videos only, thus in the main paper we train these methods only on the front cameras recording from our dataset. We breifly describe these methods as follows:

Faceformer [7]. Faceformer leverages Wav2Vec2.0 to encode audio features, which are then processed by transformer-based autoregressive model via cross-modal multi-head attention to synthesize mesh animations. The method additionally uses biased causal multi-head self attention and periodic positional encoding to improve generalization to longer sequences. The paper proposes a generic sequence model for a fixed set of identities with a style embedding to learn identity-specific speaking style.

Codetalker [21]. Given an audio signal, Codetalker formulates speech-driven facial animation as code query task of a learnt codebook. The codebook is learnt by self-reconstruction of the mesh sequences with VQ-VAE. The learnt discrete codebook is then leveraged by code-query based temporal autoregressive model for speech-conditioned facial animation. The discrete motion space of finite cardinality can accurately audio-conditioned mesh animation. Similar to Faceformer, this method also encodes audio with Wav2Vec2.0.

Imitator [20]. Similar to ours, Imitator learns a personalized model for speech-conditioned 3D facial animation. The method first pretrains a transformer based sequence model on high-quality VOCA dataset [5] and then finetunes it with short Flame tracked sequences of the personalized avatar. To model lip closures accurately, the paper further proposes a novel lip contact loss based on physiological cues of bilabial consonants 'm', 'b' 'p').

RAD-NeRF [19]. The paper proposes audio-conditioned neural radiance fields for real-time rendering. The key contribution of the method is an efficient NeRF architecture by decomposing the video representation into three low-dimensional trainable feature grids. The first two feature grid model the audio and dynamic head motion respectively. The third feature grid models the torso motion. Compared to previous audio-conditioned NeRFs, this runs much faster and enables real-time inference.

ER-NeRF [11]. Similar to RAD-NeRF, this paper also focusses on real-time rendering for audio-conditioned neural fields. However in contrast to RAD-NeRF, this method leverages triplane hash representation for learning spatial features. The method further captures the impact of audio features on different facial regions via region-aware attention module. To handle eye blinks, it uses explicit eye blinking control with a scalar. To model the torso, the method transforms a set of trainable 3D keypoints to normalized 2D coordinates and queries 2D neural field to predict the torso image.

SyncTalk [14]. Building upon ER-NeRF, this method uses triplane hash representation and further improves the quality of lip synchronization. The method leverages face-synchronization controller that align lip motion with the corresponding audio signal and uses 3D facial blendshapes for capturing facial expressions. To model head pose, it utilizes 3D head tracker and stabilizes the head pose with a pretrained optical flow estimation model. Finally, it uses a portrait-sync generator to restore rest of the details like hair and background.

GaussianTalker [3]. The paper proposes audio-conditioned talking head generation framework based on 3D Gaussian Splatting (3DGS). By leveraging the speed and efficiency of 3DGS, the authors construct a canonical 3DGS representation of the head and deform it in synchronization with the audio during audio-driven animation. Specifically, it encodes spatial information (3DGS position) of the head via multiresolution triplane feature grid and uses an MLP for predicting rest of the 3DGS attributes in canonical space. Finally, these Gaussian attributes are merged with audio features via the Spatial-Audio attention that predict per-frame 3DGS deformations, enabling stability and control.

TalkingGaussian [12]. It is a deformation-based talking head synthesis framework, also leveraging 3DGS addressing the problem of facial distortion in existing radiance field methods. The authors represent dynamic talking head with deformable 3D Gaussians, and consists of (a) static persistent Gaussians representing persistent head structure and (b) neural grid-based motion field to handle dynamic facial motion. The model is decomposed into two branches to handle face and mouth interior separately with the goal to reconstruct more accurate motion and structure of mouth region.

E. Dataset Details

Our multi-view dataset consists of native speakers in age group 20-50 and includes three male and female participants, see Tab. 4. We show additional metadata to be released with our dataset in Fig. 11 and some example frames from one of the sequences from our dataset in Fig. 12.

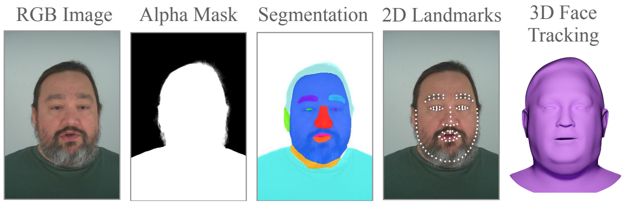


Figure 11. Dataset Metadata: Left to right, RGB Image, Alpha mask, Face segmentation, 2D landmarks and FLAME-based 3D Face Tracking. We will provide all these with our dataset release.

Participant	Native Accent	Age	Gender
Subject 1	British	27	Female
Subject 2	American	42	Male
Subject 3	American	32	Female
Subject 4	British	48	Male
Subject 5	Canadian	41	Female
Subject 6	American	32	Male

Table 4. Participant Details. We captured a gender-balanced dataset of six native English participants over a wide range of accents and age groups.

E.1. Hardware Configuration

We employ 16 machine vision cameras at a resolution of 7.1 megapixels and a supercardioid microphone to capture high-quality audio. Our capture setup is similar to Kirschstein *et al*. [10] covering a field of view of 90° left-to-right and 30° up-to-down. To avoid motion blur, we set the cameras at a shutter speed of 3ms. To capture the participant in appropriate lighting, we illuminate the subject with 8 LED light panels and use diffuser plates to reduce specularities on the skin. Our dataset can capture fine-scale facial details like eyelashes, see Fig. 13 for zoom-ins.

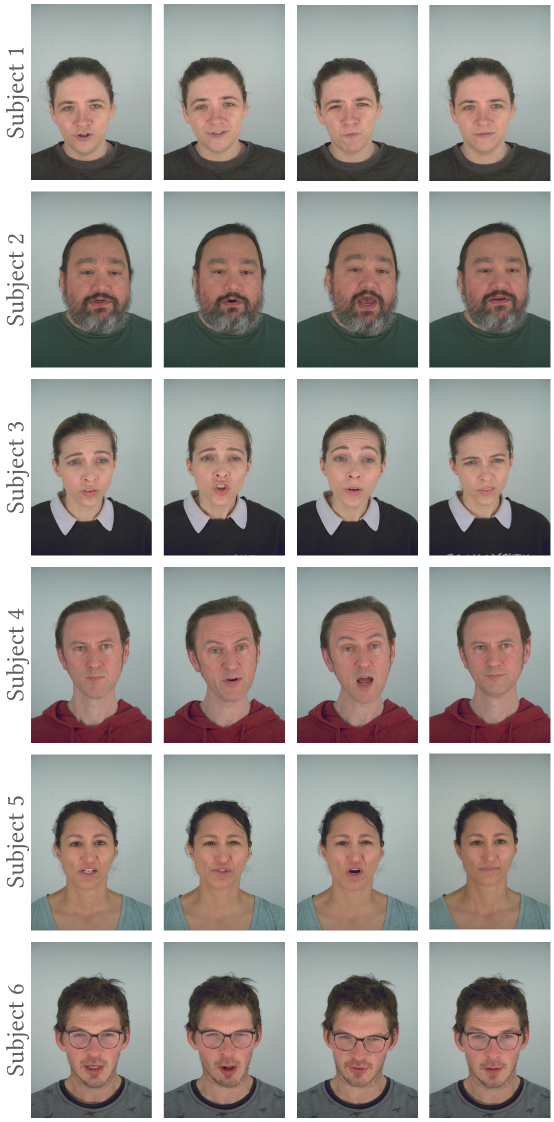


Figure 12. Dataset Participants: Four randomly selected frames from the dataset for each of six participants. The participants spoke with facial expressions and movements based on sentence transcript.

E.2. Speech Corpus

To maximize phonetic diversity in the dataset and to advance the field of audio-driven facial animation, we asked the participants to speak a phonetically diverse set of spoken English sentences carefully chosen from TIMIT speech corpus [8] with expressions. We record the following categories of sentences from TIMIT corpus (a) Two accent-

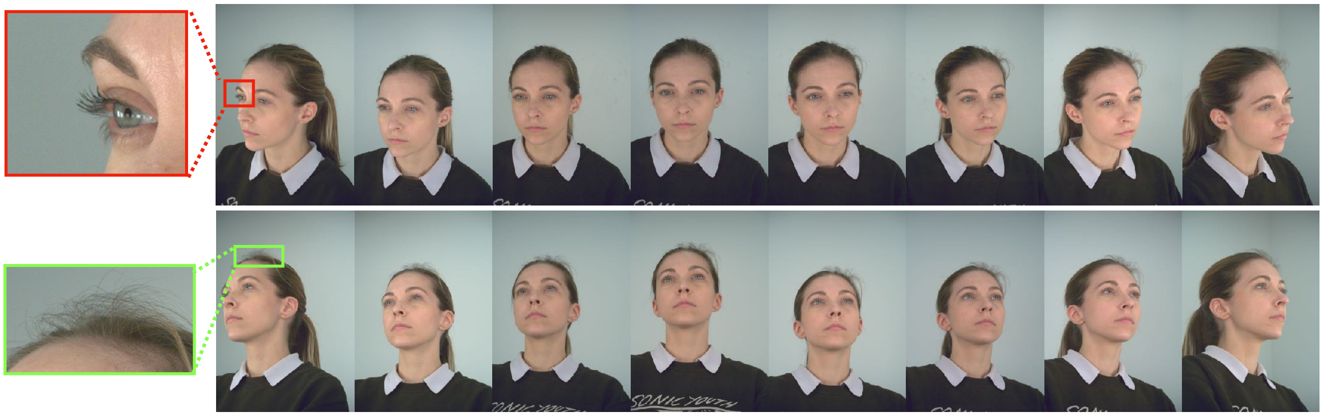


Figure 13. The multiview setup with 16 cameras used for recording participants captured at a resolution of 7.1 megapixels covering a field of view of 90° left-to-right and 30° up-to-down of the participant (right). The zoom-ins (left) for the eyes and hair show the level of detail captured by the cameras. Our audio-visual dataset contains detailed facial geometry,

specific sentences that differ for people with different dialects, (b) 260 phonetically compact sentences to include a diverse range of phonetic contexts (c) 143 phonetically balanced sentences to include a balanced representation of phonemes. These short sentences range from 3-7 seconds each. Finally, we also record 10 free-form long sentences, where we ask participants a fixed set of questions based on their hobby/profession, etc, to capture the free form speaking style of the participant. These sentences are 10-20 seconds long.

For long sequences, the participants were asked 10 basic questions as listed below.

1. Talk a little bit about your profession or education.
2. Tell us about a recent trip or vacation you took.
3. Share a hobby or activity that you enjoy pursuing in your free time.
4. Describe a cultural event or festival you attended and what made it memorable.
5. Describe a cuisine or dish you recently tried for the first time and your thoughts on it.
6. Discuss a recent technological advancement or innovation that caught your attention.
7. Talk about favourite movie/TV show.
8. Who's your favourite actor/singer?
9. What's your favourite sport?
10. Your favourite holiday destination?

For short sequences, we recorded the following sentences from TIMIT corpus.

1. She had your dark suit in greasy wash water all year.
2. Don't ask me to carry an oily rag like that.
3. Jane may earn more money by working hard.
4. Bright sunshine shimmers on the ocean.
5. Nothing is as offensive as innocence.
6. Why yell or worry over silly items?
7. Are your grades higher or lower than Nancy's?
8. Swing your arm as high as you can.
9. Before Thursday's exam, review every formula.
10. The museum hires musicians every evening.
11. Alimony harms a divorced man's wealth.
12. Aluminum silverware can often be flimsy.
13. She wore warm, fleecy, woolen overalls.
14. Those musicians harmonize marvelously.
15. Most young rise early every morning.
16. Beg that guard for one gallon of gas.
17. Help Greg to pick a peck of potatoes.
18. It's fun to roast marshmallows on a gas burner.
19. Coconut cream pie makes a nice dessert.
20. Only the most accomplished artists obtain popularity.

21. Critical equipment needs proper maintenance.
22. Young people participate in athletic activities.
23. Barb’s gold bracelet was a graduation present.
24. Stimulating discussions keep students’ attention.
25. Etiquette mandates compliance with existing regulations.
26. Biblical scholars argue history.
27. Addition and subtraction are learned skills.
28. That pickpocket was caught red-handed.
29. Grandmother outgrew her upbringing in petticoats.
30. At twilight on the twelfth day we’ll have Chablis.
31. Catastrophic economic cutbacks neglect the poor.
32. Ambidextrous pickpockets accomplish more.
33. Her classical performance gained critical acclaim.
34. Even a simple vocabulary contains symbols.
35. The eastern coast is a place for pure pleasure and excitement.
36. The lack of heat compounded the tenant’s grievances.
37. Academic aptitude guarantees your diploma.
38. The prowler wore a ski mask for disguise.
39. We experience distress and frustration obtaining our degrees.
40. The legislature met to judge the state of public education.
41. Chocolate and roses never fail as a romantic gift.
42. Any contributions will be greatly appreciated.
43. Continental drift is a geological theory.
44. We got drenched from the uninterrupted rain.
45. Last year’s gas shortage caused steep price increases.
46. Upgrade your status to reflect your wealth.
47. Eat your raisins outdoors on the porch steps.
48. Porcupines resemble sea urchins.
49. Cliff’s display was misplaced on the screen.
50. An official deadline cannot be postponed.
51. Fill that canteen with fresh spring water.
52. Gently place Jim’s foam sculpture in the box.
53. Bagpipes and bongos are musical instruments.
54. Doctors prescribe drugs too freely.
55. Will you please describe the idiotic predicament.
56. It’s impossible to deal with bureaucracy.
57. Good service should be rewarded by big tips.
58. My instructions desperately need updating.
59. Cooperation along with understanding alleviate dispute.
60. Primitive tribes have an upbeat attitude.
61. Flying standby can be practical if you want to save money.
62. The misprint provoked an immediate disclaimer.
63. A large household needs lots of appliances.
64. Youngsters love common candy as treats.
65. Iguanas and alligators are tropical reptiles.
66. Masquerade parties tax one’s imagination.
67. Penguins live near the icy Antarctic.
68. Medieval society was based on hierarchies.
69. Project development was proceeding too slowly.
70. Kindergarten children decorate their classrooms for all holidays.
71. Special task forces rescue hostages from kidnappers.
72. Call an ambulance for medical assistance.
73. He stole a dime from a beggar.
74. A huge tapestry hung in her hallway.
75. Birthday parties have cupcakes and ice cream.
76. His scalp was blistered from today’s hot sun.

77. She slipped and sprained her ankle on the steep slope.
78. The best way to learn is to solve extra problems.
79. Tugboats are capable of hauling huge loads.
80. A muscular abdomen is good for your back.
81. The cartoon features a muskrat and a tadpole.
82. The emblem depicts the Acropolis all aglow.
83. The mango and the papaya are in a bowl.
84. Combine all the ingredients in a large bowl.
85. The misquote was retracted with an apology.
86. The coyote, bobcat, and hyena are wild animals.
87. Trespassing is forbidden and subject to penalty.
88. Encyclopedias seldom present anecdotal evidence.
89. A screwdriver is made from vodka and orange juice.
90. Westchester is a county in New York.
91. Artificial intelligence is for real.
92. Lots of foreign movies have subtitles.
93. Angora cats are furrier than Siamese.
94. Publicity and notoriety go hand in hand.
95. Pizzerias are convenient for a quick lunch.
96. December and January are nice months to spend in Miami.
97. Technical writers can abbreviate in bibliographies.
98. Scientific progress comes from the development of new techniques.
99. Tradition requires parental approval for under-age marriage.
100. The clumsy customer spilled some expensive perfume.
101. The bungalow was pleasantly situated near the shore.
102. Pledge to participate in Nevada’s aquatic competition.
103. Which long article was opaque and needed clarification?
104. The sound of Jennifer’s bugle scared the antelope.
105. The willowy woman wore a muskrat coat.
106. Too much curiosity can get you into trouble.
107. Correct execution of my instructions is crucial.
108. Most precincts had a third of the votes counted.
109. While waiting for Chipper she crisscrossed the square many times.
110. The previous speaker presented ambiguous results.
111. Mosquitoes exist in warm, humid climates.
112. Scholastic aptitude is judged by standardized tests.
113. Orange juice tastes funny after toothpaste.
114. The water contained too much chlorine and stung his eyes.
115. Our experiment’s positive outcome was unexpected.
116. Remove the splinter with a pair of tweezers.
117. The government sought authorization of his citizenship.
118. As coauthors, we presented our new book to the haughty audience.
119. As a precaution, the outlaws bought gunpowder for their stronghold.
120. Her auburn hair reminded him of autumn leaves.
121. They remained lifelong friends and companions.
122. Curiosity and mediocrity seldom coexist.
123. The easygoing zoologist relaxed throughout the voyage.
124. Biologists use radioactive isotopes to study microorganisms.
125. Employee layoffs coincided with the company’s reorganization.
126. How would you evaluate this algebraic expression?
127. The Mayan neoclassic scholar disappeared while surveying ancient ruins.
128. The diagnosis was discouraging; however, he was not overly worried.
129. The triumphant warrior exhibited naive heroism.
130. Whoever cooperates in finding Nan’s cameo will be rewarded.

131. The haunted house was a hit due to outstanding audio-visual effects.
132. Severe myopia contributed to Ron's inferiority complex.
133. Buying a thoroughbred horse requires intuition and expertise.
134. She encouraged her children to make their own Halloween costumes.
135. We could barely see the fjords through the snow flurries.
136. Almost all colleges are now coeducational.
137. Rich looked for spotted hyenas and jaguars on the safari.
138. Why else would Danny allow others to go?
139. Who authorized the unlimited expense account?
140. Destroy every file related to my audits.
141. Serve the coleslaw after I add the oil.
142. Withdraw all phony accusations at once.
143. Straw hats are out of fashion this year.
144. Draw each graph on a new axis.
145. Norwegian sweaters are made of lamb's wool.
146. Young children should avoid exposure to contagious diseases.
147. Ralph controlled the stopwatch from the bleachers.
148. Approach your interview with statuesque composure.
149. The causeway ended abruptly at the shore.
150. Even I occasionally get the Monday blues!
151. Military personnel are expected to obey government orders.
152. When peeling an orange, it is hard not to spray juice.
153. Rob sat by the pond and sketched the stray geese.
154. Michael colored the bedroom wall with crayons.
155. I gave them several choices and let them set the priorities.
156. The news agency hired a great journalist.
157. The morning dew on the spider web glistened in the sun.
158. The sermon emphasized the need for affirmative action.
159. The small boy put the worm on the hook.
160. Try to recall the events in chronological order.
161. Nonprofit organizations have frequent fund raisers.
162. The most recent geological survey found seismic activity.
163. Cory attacked the project with extra determination.
164. You always come up with pathological examples.
165. Put the butcher block table in the garage.
166. Keep the thermometer under your tongue!
167. Steph could barely handle the psychological trauma.
168. It's healthier to cook without sugar.
169. Allow leeway here, but rationalize all errors.
170. His failure to open the store by eight cost him his job.
171. Highway and freeway mean the same thing.
172. The paper boy bought two apples and three ices.
173. Clear pronunciation is appreciated.
174. A doctor was in the ambulance with the patient.
175. Puree some fruit before preparing the skewers.
176. It's not easy to create illuminating examples.
177. The hallway opens into a huge chamber.
178. May I order a strawberry sundae after I eat dinner?
179. They all agree that the essay is barely intelligible.
180. Herb's birthday occurs frequently on Thanksgiving.
181. The cigarettes in the clay ashtray overflowed onto the oak table.
182. Reading in poor light gives you eyestrain.
183. The Boston Ballet overcame their funding shortage.
184. We apply auditory modeling to computer speech recognition.
185. The gorgeous butterfly ate a lot of

nectar.
186. Tornados often destroy acres of farm land.
187. Remember to allow identical twins to enter freely.
188. How oily do you like your salad dressing?
189. We saw eight tiny icicles below our roof.
190. The saw is broken, so chop the wood instead.
191. Withdraw only as much money as you need.
192. Draw every outer line first, then fill in the interior.
193. The jaw operates by using antagonistic muscles.
194. Cliff was soothed by the luxurious massage.
195. Steve wore a bright red cashmere sweater.
196. To further his prestige, he occasionally reads the Wall Street Journal.
197. Alice's ability to work without supervision is noteworthy.
198. Cory and Trish played tag with beach balls for hours.
199. The tooth fairy forgot to come when Roger's tooth fell out.
200. Planned parenthood organizations promote birth control.
201. Jeff thought you argued in favor of a centrifuge purchase.
202. Rich purchased several signed lithographs.
203. In every major cloverleaf, traffic sometimes gets backed up.
204. In the long run, it pays to buy quality clothing.
205. Brush fires are common in the dry underbrush of Nevada.
206. Weatherproof galoshes are very useful in Seattle.
207. This brochure is particularly informative for a prospective buyer.
208. The avalanche triggered a minor earthquake.
209. These exclusive documents must be locked up at all times.
210. Please take this dirty table cloth to the cleaners for me.
211. Should giraffes be kept in small zoos?
212. If Carol comes tomorrow, have her arrange for a meeting at two.
213. I'd rather not buy these shoes than be overcharged.
214. Shaving cream is a popular item on Halloween.
215. Amoebas change shape constantly.
216. We like bleu cheese but Victor prefers swiss cheese.
217. Tofu is made from processed soybeans.
218. The bluejay flew over the high building.
219. Cheap stockings run the first time they're worn.
220. Cottage cheese with chives is delicious.
221. Shipbuilding is a most fascinating process.
222. The proof that you are seeking is not available in books.
223. The hood of the jeep was steaming in the hot sun.
224. My desires are simple: give me one informative paragraph on the subject.
225. Those answers will be straightforward if you think them through carefully first.
226. If people were more generous, there would be no need for welfare.
227. The nearest synagogue may not be within walking distance.
228. The groundhog clearly saw his shadow, but stayed out only a moment.
229. The local drugstore was charged with illegally dispensing tranquilizers.
230. Al received a joint appointment in the biology and the engineering departments.
231. Gregory and Tom chose to watch cartoons in the afternoon.
232. Chip postponed alimony payments until the latest possible date.
233. Count the number of teaspoons of soysauce that you add.
234. The big dog loved to chew on the old rag doll.
235. Todd placed top priority on getting his bike fixed.
236. An adult male baboon's teeth are not suitable for eating shellfish.
237. Often you'll get back more than you put in.

238. Gus saw pine trees and redwoods on his walk through Sequoia National Forest.
239. Rob made Hungarian goulash for dinner and gooseberry pie for dessert.
240. Bob bandaged both wounds with the skill of a doctor.
241. The high security prison was surrounded by barbed wire.
242. Take charge of choosing her bride's maids' gowns.
243. The frightened child was gently subdued by his big brother.
244. The barracuda recoiled from the serpent's poisonous fangs.
245. The patient and the surgeon are both recuperating from the lengthy operation.
246. I'll have a scoop of that exotic purple and turquoise sherbet.
247. The preschooler couldn't verbalize her feelings about the emergency conditions.
248. Many wealthy tycoons splurged and bought both a yacht and a schooner.
249. The new suburbanites worked hard on refurbishing their older home.
250. According to my interpretation of the problem, two lines must be perpendicular.
251. The system may break down soon, so save your files frequently.
252. The annoying raccoons slipped into Phil's garden every night.
253. I took her word for it, but is she really going with you?
254. The gunman kept his victim cornered at gunpoint for three hours.
255. Will you please confirm government policy regarding waste removal?
256. The fish began to leap frantically on the surface of the small lake.
257. Her wardrobe consists of only skirts and blouses.
258. There was a gigantic wasp next to Irving's big top hat.
259. Those who are not purists use canned vegetables when making stew.
260. They used an aggressive policeman to flag thoughtless motorists.
261. Shell shock caused by shrapnel is sometimes cured through group therapy.
262. Ralph prepared red snapper with fresh lemon sauce for dinner.
263. If you destroy confidence in banks, you do something to the economy, he said.
264. He further proposed grants of an unspecified sum for experimental hospitals.
265. Nothing has been done yet to take advantage of the enabling legislation.
266. It also provides for funds to clear slums and help colleges build dormitories.
267. The prospect of cutting back spending is an unpleasant one for any governor.
268. He really crucified him; he nailed it for a yard loss.
269. There is definitely some ligament damage in his knee.
270. In fact our whole defensive unit did a good job.
271. He played basketball there while working toward a law degree.
272. So, if anybody solicits by phone, make sure you mail the dough to the above.
273. Her position covers a number of daily tasks common to any social director.
274. The structures housing the apartments are of masonry and frame construction.
275. This, he added, brought about petty jealousies and petty personal grievances.
276. There was no confirmation of such massive assaults from independent sources.
277. The staff deserves a lot of credit working down here under real obstacles.
278. They make gin saws and deal in parts, supplies and some used gin machinery.
279. Maybe it's taking longer to get things squared away than the bankers expected.
280. Hiring the wife for one's company may win her tax-aided retirement income.
281. Unfortunately, there is still little demand for broccoli and cauliflower.
282. Displayed as lamps, the puppets delight the children and are

decorative accent.

283. To create such a lamp, order a wired pedestal from any lamp shop.

284. There are more obvious nymphomaniacs on any private-eye series.

285. But this doesn't detract from its merit as an interesting, if not great, film.

286. And you think you have language problems.

287. Ideally, he knew, it should be preceded by concrete progress at lower levels.

288. This is a significant advance but its import should not be exaggerated.

289. Adequate compensation is indispensable.

290. This is a problem that goes considerably beyond questions of salary and tenure.

291. Some observers speculated that this might be his revenge on his home town.

292. Confusion became chaos; each succeeding day brought new acts of violence.

293. That added traffic means rising streams of dimes and quarters at toll gates.

294. Traffic frequently has failed to measure up to engineers' rosy estimates.

295. Progress is being made, too, in improving motorists' access to many turnpikes.

296. Under this law annual grants are given to systems in substantial amounts.

297. Within a system, however, the autonomy of each member library is preserved.

298. The desire and ability to read are important aspects of our cultural life.

299. We congratulate the entire membership on its record of good legislation.

300. Thereupon followed a demonstration that tyranny knows no ideological confines.

301. Wooded stream valleys in the folds of earth would be saved.

302. His election, on the other hand, would unquestionably strengthen the regulars.

303. He spoke briefly, sensibly, to the point and without oratorical flourishes.

304. Further, it has its work cut out stopping anarchy where it is now garrisoned.

305. Fools, he bayed, what do you think you are doing?

306. So we note approvingly a fresh sample of unanimity.

307. It is one of the rare public ventures here on which nearly everyone is agreed.

308. Thus there is a clearer division of authority, administrative and legislative.

309. Jokes, cartoons and cynics to the contrary, mothers-in-law make good friends.

310. Theirs is a sacrificial life by earthly standards.

311. The narrow fringe of sadness that ran around it only emphasized the pleasure.

312. Would a blue feather in a man's hat make him happy all day?

313. These programs emphasize the acceptance of biracial classrooms peacefully.

314. You certainly can't expect the infield to do any better than it did last year.

315. Is the mother of an autistic child at fault?

316. As a rule, the autistic child doesn't enjoy physical contact with others.

317. Or certain words or rituals that child and adult go through may do the trick.

318. We did not accept the diagnosis at once, but gradually we are coming to.

319. Is a relaxed home atmosphere enough to help her outgrow these traits?

320. Where only one club existed before, he says, two will flourish henceforth.

321. This is going to be a language lesson, and you can master it in a few minutes.

322. Family loyalties and cooperative work have been unbroken for generations.

323. Heels place emphasis on the long

legged silhouette.
324. Wine glass heels are to be found in both high and semi-heights.
325. Stacked heels are also popular on dressy or tailored shoes.
326. Contrast trim provides other touches of color.
327. At the left is a pair of dressy straw pumps in a light, but crisp texture.
328. At right is a casual style in a crushed unlined white leather.
329. Most of us brush our teeth by hand.
330. The bristles are soft enough to massage the gums and not scratch the enamel.
331. "Steam baths" writes: do steam baths have any health value?
332. "Sewing brings numbness" writes: what makes my hands numb when sewing?
333. Teaching guides are included with each record.
334. He doesn't want her to look frowningly at him, or speak to him angrily.
335. But even mother's loving attitude will not always prevent misbehavior.
336. She can decrease the number of temptations.
337. She can remove all knick-knacks within reach.
338. Usually, they titter loudly after they have passed by.
339. Too often, unless he hails them, they pass him by.
340. Say he is a horse thief, runs an old adage.
341. It seems that open season upon veterans' hospitalization is once more upon us.
342. This we can sympathetically understand.
343. This is taxation without representation.
344. Our entire economy will have a terrific uplift.
345. One even gave my little dog a biscuit.
346. Maybe he will help to turn our fair city into a ghost town.
347. Why do we need bigger and better bombs?
348. One of the problems associated with the expressway stems from the basic idea.
349. Bridges, tunnels and ferries are the most common methods of river crossings.
350. Replace it with the statue of one or another of the world's famous dictators.
351. The gallant half-city is dying on its feet.
352. Their privations are almost beyond endurance.
353. The moment of truth is the moment of crisis.
354. New self-deceiving rags are hurriedly tossed on the too-naked bones.
355. What explains this uni-directional paralysis?
356. Originals are not necessarily good and adaptations are not necessarily bad.
357. But that explanation is only partly true.
358. But the ships are very slow now, and we don't get so many sailors any more.
359. They were shown how to advance against an enemy outpost atop a cleared ridge.
360. We would lose our export markets and deny ourselves the imports we need.
361. With this no loyal citizen can quarrel.
362. This possibility is anything but reassuring.
363. The public is now armed with sophistication and numerous competing media.
364. But the attack was made from an advance copy.
365. Well, now we have two big theaters.
366. Splendor by sorcery: it's a horror.
367. One-upmanship is practiced by both sides in a total war.
368. This big, flexible voice with uncommon range has been superbly disciplined.
369. Her debut over, perhaps the earlier scenes will emerge equally fine.
370. He injected more vitality into the score than it has revealed in many years.
371. The storyline, in sort, is wildly unrealistic.

372. He talked about unauthentic storylines too.
373. He praises many individuals generously.
374. His portrayal of an edgy head-in-the-clouds artist is virtually flawless.
375. Not a corner has been visibly cut in this one.
376. The master's hand has lost none of its craft.
377. He showed puny men attacked by splendidly tyrannical machines.
378. He may have a point in urging that decadent themes be given fewer prizes.
379. The works are presented chronologically.
380. The humor of the situation can be imagined.
381. His technique is ample and his musical ideas are projected beautifully.
382. What a discussion can ensue when the title of this type of song is in question.
383. Program note reads as follows: take hands; this urgent visage beckons us.
384. The orchestra was obviously on its mettle and it played most responsively.
385. He liked to nip ear lobes of unsuspecting visitors with his needle-sharp teeth.
386. Here, he is, quite persuasively, the very embodiment of meanness and slyness.
387. He is a man of major talent -- but a man of solitary, uncertain impulses.
388. He was above all a friend seeker, almost pathetic in his eagerness to be liked.
389. He enlisted a staff of loyal experts and of many zealous volunteers.
390. But what has been happening recently might be described as creeping mannerism.
391. Clever light songs were overly coy, tragic songs a little too melodramatic.
392. Below is a specific guide, keyed to the calendar.
393. A sailboat may have a bone in her teeth one minute and lie becalmed the next.
394. It suffers from a lack of unity of purpose and respect for heroic leadership.
395. The fat man has trouble buying life insurance or has to pay higher premiums.
396. Far more frequently, overeating is the result of a psychological compulsion.
397. Yet it exists and has an objective reality which can be experienced and known.
398. Yet the spirit which lives in community is not identical with the community.
399. But this statement is completely unconvincing.
400. A second point requires more extended comment.
401. The straight line would symbolize its uniqueness, the circle its universality.
402. His history is his alone, yet each man must recognize his own history in it.
403. Death reminds man of his sin, but it reminds him also of his transience.
404. Such a calm and assuring peace can be yours.
405. Satellites, sputniks, rockets, balloons; what next?

References

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

[2] Lukas Biewald. Experiment tracking with weights and biases, 2020. Software available from wandb.com. 4

[3] 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. 4, 6

[4] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016. 5

[5] 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. 6

[6] William Falcon et al. Pytorch lightning. *GitHub. Note: https://github. com/PyTorchLightning/pytorch-lightning*, 3 (6), 2019. 4

[7] 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. 4, 6

[8] 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. 7

[9] 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. 4, 5, 6

[10] 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. 7

[11] 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. 4, 6

[12] 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. 3, 4, 7

[13] 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. 1, 5

[14] 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. 4, 6

[15] 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. 5

[16] 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. 1, 4, 5, 6

[17] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, page 497–500, New York, NY, USA, 2001. Association for Computing Machinery. 5

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

[19] 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. 4, 6

[20] 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. 4, 6

[21] 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. 4, 6