

# GaussianTalker: Real-Time High-Fidelity Talking Head Synthesis with Audio-Driven 3D Gaussian Splatting - Supplementary Materials -

Anonymous Authors

In the supplementary document, we describe the implementation details and further analyses of **GaussianTalker**. Specifically, we first introduce the details of our network design and hyperparameter settings in Sec. A. We also provide details of our analysis on the proposed method that was conducted in the main paper in Sec. B. In Sec. C, we validate our methodology with more qualitative results from our experiments, and also conduct a user study. Then, more ablation studies are conducted in Sec. D. To further demonstrate the robustness and effectiveness of our framework, we also provide a supplementary video (Sec. E). Finally, we discuss the limitations and ethical considerations of our research in Sec. F.

## A IMPLEMENTATION DETAILS

### A.1 Network architecture

**A.1.1 Multi-resolution Triplane.** Our multi-resolution triplane consists of three orthogonal grids, with the hidden feature dimension of  $H = 64$ , and its base resolution of  $R = 64$ , which is further upsampled by 2.

**A.1.2 Canonical 3D Gaussian attribute predictor.** The employed network that predicts the attributes of canonical 3D Gaussians is made up of MLPs, such as:  $\mathcal{F}_{\text{can}} = \{\phi_{\text{shared}}, \phi_r, \phi_s, \phi_{SH}, \phi_\alpha\}$ . Specifically, a tiny MLP  $\phi_{\text{shared}}$  encodes the triplane embedding  $f(\mu_c)$  and outputs a shared feature  $\kappa$  for all attributes. The following MLP regressors maps this feature to each 3D Gaussian attribute such as:

$$\begin{aligned} \kappa &= \phi_{\text{shared}}(f(\mu)), \\ r_c &= \phi_r(\kappa), s_c = \phi_s(\kappa), SH_c = \phi_{SH}(\kappa), \alpha_c = \phi_\alpha(\kappa). \end{aligned} \quad (\text{A1})$$

**A.1.3 Deformation offset predictor.** Similar to  $\mathcal{F}_{\text{can}}$ , the deformation prediction network,  $\mathcal{F}_{\text{can}} = \{\psi_\mu, \psi_r, \psi_s, \psi_{SH}, \psi_\alpha\}$ , that estimates the deformation offsets of each Gaussian attribute for each frame consists of several small MLP regressors. For the  $n$ -th frame, the final output embedding from the cross-attention module,  $z_n^L$ , is mapped to each attribute offset such that

$$\begin{aligned} \Delta\mu_n &= \psi_\mu(z_n^L), \Delta r_n = \psi_r(z_n^L), \Delta s_n = \psi_s(z_n^L), \\ \Delta SH_n &= \psi_{SH}(z_n^L), \Delta\alpha_n = \psi_\alpha(z_n^L). \end{aligned} \quad (\text{A2})$$

### A.2 Hyperparameter Configuration

During the **canonical stage**, we conduct training over 8,000 iterations for a specific identity. We set the weights for the loss functions as follows:  $\lambda_1 = 0.8$ ,  $\lambda_{\text{lpips}} = 0.01$ , and  $\lambda_{\text{D-SSIM}} = 0.2$ . The initial learning rate for the multi-resolution triplane is set to 0.0016, gradually decaying to 0.00016. Similarly, the learning rate for  $\mathcal{F}_{\text{can}}$  starts at 0.0001 and diminishes to 0.00001. We cap the maximum number of 3D Gaussians at 50,000, and we abstain from utilizing the opacity reset operation from the original implementation [3], as we found it does not yield discernible benefits in our experiments.

Subsequently, in the **deformation stage**, we proceed with training the network for 8,000 iterations. We maintain the same weighting scheme for the loss functions:  $\lambda_1 = 0.8$ ,  $\lambda_{\text{lpips}} = 0.01$ ,  $\lambda_{\text{D-SSIM}} = 0.2$ , and  $\lambda_{\text{lip}} = 0.8$ . All modules are trained with an initial learning rate of 0.0001, gradually decreasing to 0.00001.

While our spatial-audio cross-attention module primarily employs  $L = 2$  cross-attention layers, our modified GaussianTalker\* with  $L = 1$  can achieve comparable results with even faster inference speeds.

### A.3 Splatting on the background image

Initially, our research followed the method outlined in the original implementation [3], where faces were generated on a white background. However, we encountered limitations with this approach. To render images containing only faces on a white background, corresponding ground truth images with similar characteristics were required, necessitating the use of a segmentation model. However, due to the inherent inaccuracies of the segmentation model, the obtained facial masks tended to encompass larger areas, including the background. Additionally, the disproportionate emphasis of loss terms such as SSIM and perceptual loss on imperfect facial contours relative to mouth and eye movements hindered the learning process.

As a solution, we opted to generate faces against GT backgrounds instead. This approach allowed for the accurate learning of Gaussian presence boundaries by distributing loss across the entire image. Similar to preprocessing techniques employed in previous NeRF-based works [2, 4, 6], we interpolated the human form from the background image to create an image with the person removed. Subsequently, faces were directly rendered using Gaussian methods, enabling comparisons with GT videos. By adopting this strategy, our GaussianTalker is trained without the need for facial mask, facilitating the faithful representation of intricate details such as hair.

## B DETAILED ANALYSIS AND VISUALIZATION

### B.1 Analysis of attention

We demonstrate the effectiveness of our **spatial-audio cross attention module**, by presenting more visualization of attention scores for different input conditions. These visualizations highlight which input features most influence the deformations of specific 3D Gaussians. As observed, speech audio attention scores are primarily concentrated around the lip and mouth regions, indicating their dominance in controlling lip movements. In contrast, the eye blink feature  $e$  focuses its attention on the eye region, as expected. The viewpoint condition  $v$  represents head orientation and influences facial shadows and wrinkles distributed across the entire head.

## B.2 Visualization of Attention

In our spatial-audio cross-attention module, the computation of the attention score is formalized by the following equation:

$$(A_n)^l = \frac{\text{softmax}(qk_n^T)^l}{\sqrt{d_k}}, \quad (\text{A3})$$

where  $l$  denotes the index of  $\{a_n, e_n, v_n, \emptyset\}$  and  $(A_n)^l$  corresponds to its calculated attention score.  $A_n$  denotes the concatenation of all  $(A_n)^l$ , resulting in a shape of  $B \times H \times N \times d_k$ , which respectively indicate batch size, number of heads, number of Gaussians, and number of features per Gaussian.

For each attention score  $(A_n)^l$ , we visualize the attention by assigning the score to RGB values. Thereby we obtain attention visualization colors  $c_{att}$  for each Gaussian. The overall visualization of attention is then calculated such as:

$$C = \sum_{i=1} c_i \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j), \quad (\text{A4})$$

where  $c_i$  represents the color associated with each Gaussian, determined by  $c_{att}$  along the view direction.  $\alpha'_i$  is derived from the multiplication of the opacity  $\alpha$  of the 3D Gaussian and its projected covariance  $\Sigma'$ . This mathematical formulation allows us to visually interpret the model's focus within the generated representations, effectively highlighting the areas of greatest feature impact.

## B.3 Visualization of triplane

Fig. 3 of the main paper visualizes the PCA analysis result of our multi-resolution triplane, showing the efficacy of using triplane to embed Gaussian features. We perform PCA on each triplane with dimensions  $H \times R \times R$ , linearly transforming the first dimension down to three principal components, resulting in dimensions  $3 \times R \times R$ . Subsequently, the values of the first dimension are normalized between  $[0, 255]$  to denote RGB values. As a result, in all xy, yz, and zx triplanes, semantically close facial regions are consistently represented with similar colorations.

## C ADDITIONAL EXPERIMENTS

### C.1 Additional qualitative experiments.

We present additional attention map visualization on Fig. A1, and also present additional visualization of generated keyframes from comparison experiments in the **self-driven** setting and the **cross-driven** setting in Fig. A2 and Fig. A3 respectively. These experiments showcase the stability of our method and its applicability to various identities.

### C.2 User study

Following previous works [4, 6], we conducted a user study in order to better judge the visual quality of the generated talking head videos. 21 participants with an age range of 20-40 years old were solicited to evaluate the rendered results in the head reconstruction setting. For accurate judgments, we combine all generated videos into a single high-resolution video, enabling simultaneous observation of all movements by the participants. To ensure fairness in the comparison process, we assign a number to each generated result instead of identifying them by their method. Participants were

asked to evaluate the three perspectives of the generated portraits: (1) Lip-sync Accuracy; (2) Video Realness; and (3) Image Quality. The results are shown in Tab. A1.

175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232

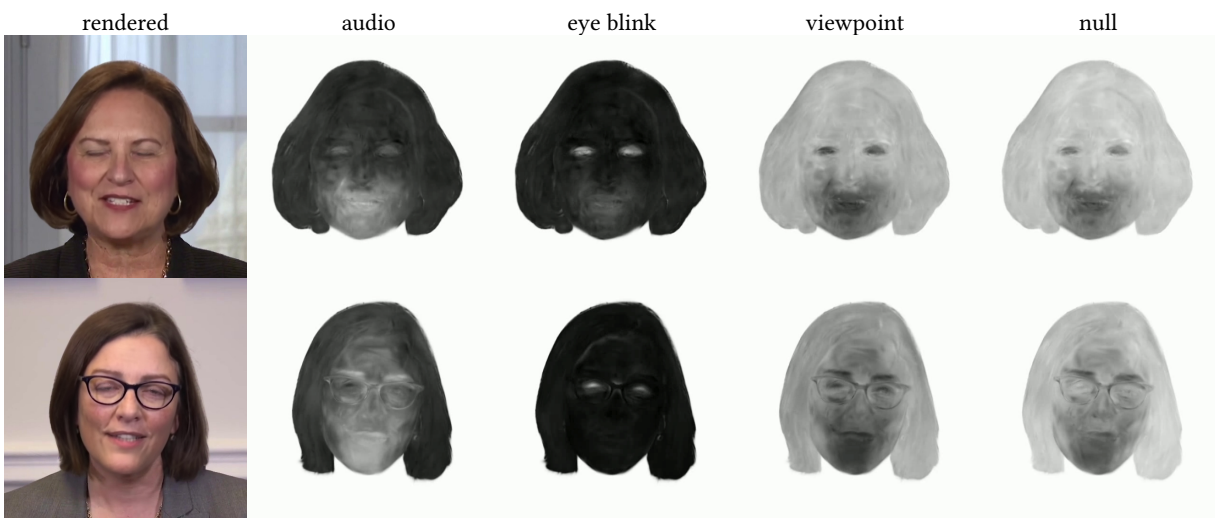


Figure A1: More results comparison on the *self-driven* setting.

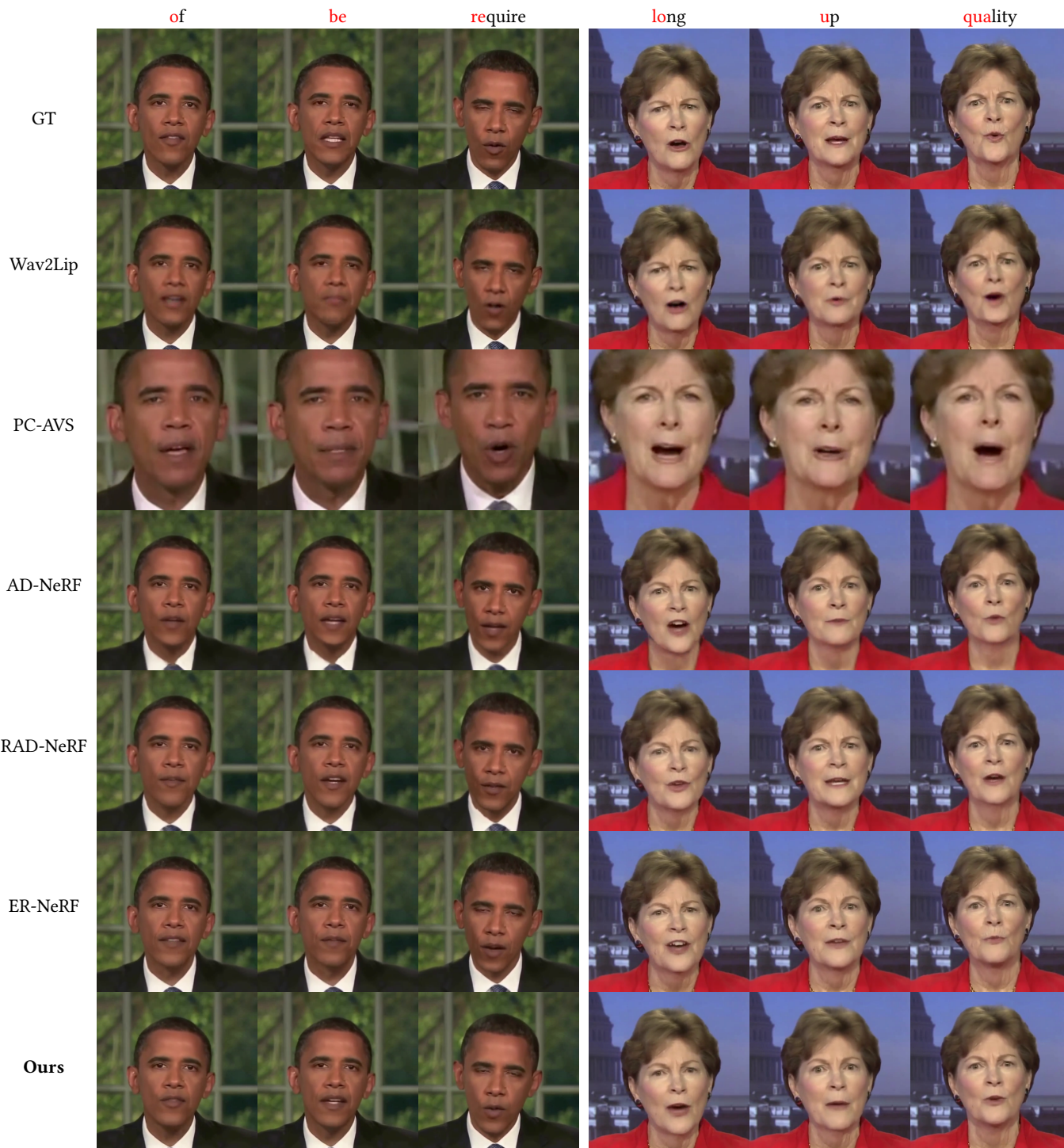
Figure A2: More results comparison on the *self-driven* setting.





Figure A3: More results comparison on the *cross-driven* setting.

Methods	self-driven			cross-driven		
	Lip-sync Accuracy	Image Quality	Video Realness	Lip-sync Accuracy	Image Quality	Video Realness
Wav2Lip [5]	3.167	2.665	2.459	2.678	2.313	2.135
PC-AVS [7]	2.625	1.896	1.921	1.958	1.292	1.229
AD-NeRF [2]	2.031	2.492	2.396	2.574	3.042	2.365
RAD-NeRF [6]	2.417	2.750	2.541	2.938	3.146	2.604
ER-NeRF [4]	2.354	3.042	2.771	2.792	3.458	3.146
GaussianTalker	3.083	3.667	3.188	3.250	3.729	3.208

Table A1: User study results. The rating is of scale 1-5, the higher the better. The top, second-best, and third-best results are shown in red, orange, and yellow, respectively.

## D ABLATION STUDIES

### D.1 Initialization of $\mu_c$

Our study explores the impact of initialization on canonical 3D Gaussian optimization. In the default setting, we leverage a pre-optimized Basel Face Model [1] to obtain camera parameters during preprocessing. These optimized mesh vertices are used to initialize the 3D positions,  $\mu_c$  of the 3D Gaussians.

To investigate the impact of the proposed 3DMM-based initialization, we conduct an ablation study by comparing it to random initialization from a sphere. In Fig. A4, we visually analyze the optimization process of the canonical stage under both initialization settings. Our experiments demonstrate that utilizing 3DMM-based initialization leads to faster convergence, attributed to the facial depth information encoded in the initialized points.

### D.2 Selection of attributes inferred for triplane embeddings.

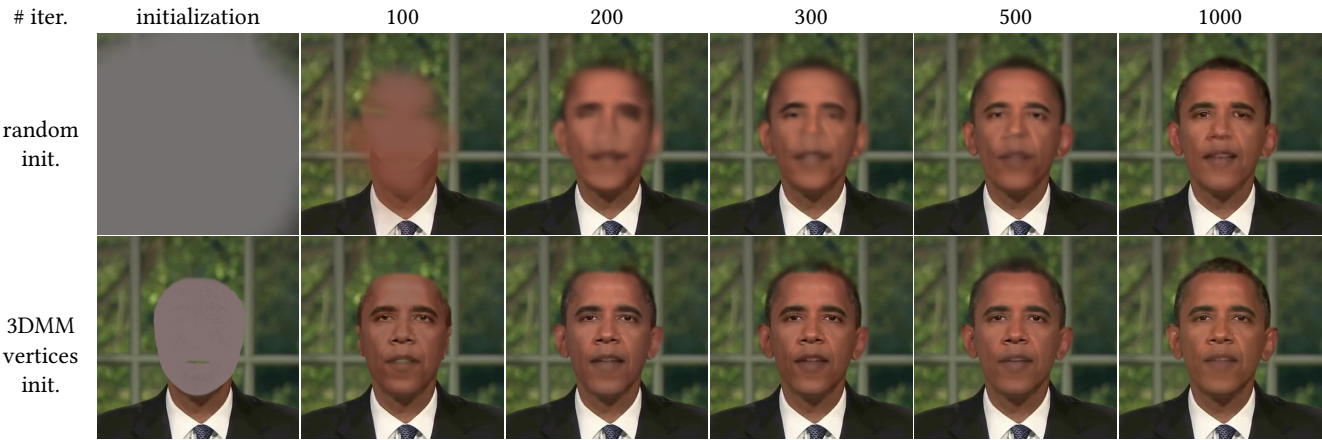
In Fig. A5, we support the quantitative comparison in the main paper by presenting key frames of the rendered results. Conditioning the triplane embeddings on the structure information such as  $r$  and  $s$  tends to show less accurate facial details such as wrinkles in facial muscle. In contrast, while conditioning on appearance information  $SH$  and  $\alpha$  produce accurate reconstructions of the canonical head, the facial motion appears less dynamic compared to the ground truth, and does not correlate well with input speech audio.

### D.3 Selection of deformed attributes.

We also provide qualitative comparisons from our ablation study on selection of Gaussian attributes to be deformed. Utilizing the same comparison settings from Sec.5.4.2, we visualize the rendered results in Fig. A6. Only deforming  $SH$  and  $\alpha$  show blurry results with unrealistic deformations, while only manipulating

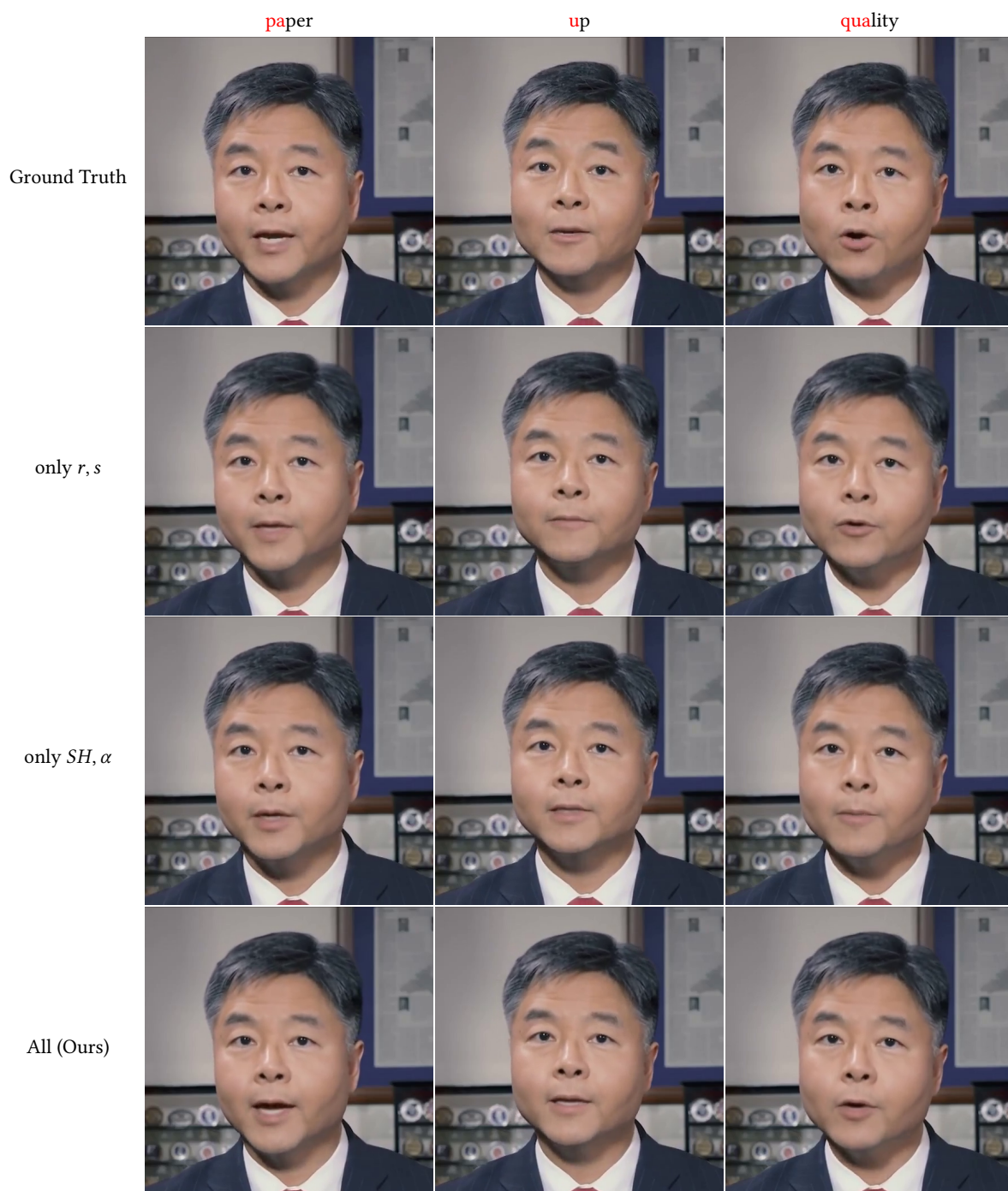
### D.4 Disentanglement of audio-unrelated motion

Finally, we reinforce the insights drawn from the quantitative analysis in Section 5.4.3. We elucidate the disentanglement of speech-related motion in Fig. A7 by presenting visualizations of the attention scores for the input conditions across the ablation experiment settings. Notably, the attention scores of the input speech audio become more widely distributed across other facial regions, indicating inadequate disentanglement of speech-related motion when solely provided with speech as the input condition.



**Figure A4: Ablation study on initialization of the canonical position  $\mu_c$ .** We evaluate the effectiveness of the 3DMM-based initialization by visualizing the optimization process of the reconstructed canonical 3D head, and compare it to random initialization. Our experiments demonstrate that utilizing 3DMM-based initialization leverages the depth information of the human face, leading to significantly faster convergence. In contrast, optimizing from randomly sampled points prolongs training duration and fails to completely resolve artifacts, particularly around the eyes and hair regions.





**Figure A5: Ablation study on the selection of attributes inferred from the triplane embedding  $f(\mu)$ . We compare the generated results from**

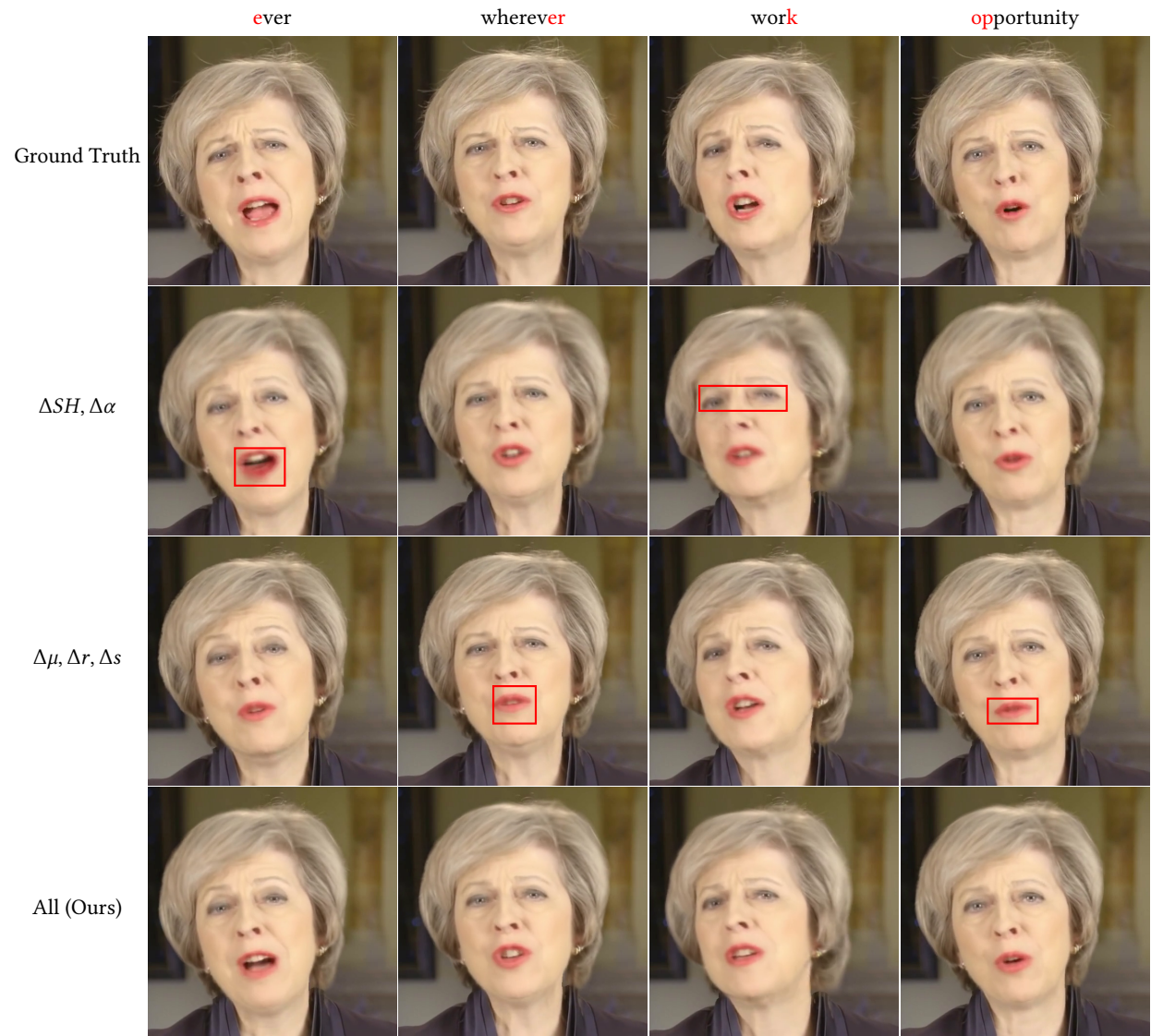


Figure A6: Deforming only spherical harmonics and opacity resulted in a significant loss of facial detail and blurry reconstructions. Notably, this led to unrealistic deformations in lip regions, where the lips and teeth appeared merged. Conversely, deforming only structural information ( $\mu, r, s$ ) produced much less dynamic lip movements. In addition, the generated results show the inside of the mouth, such as teeth and tongue less frequently.

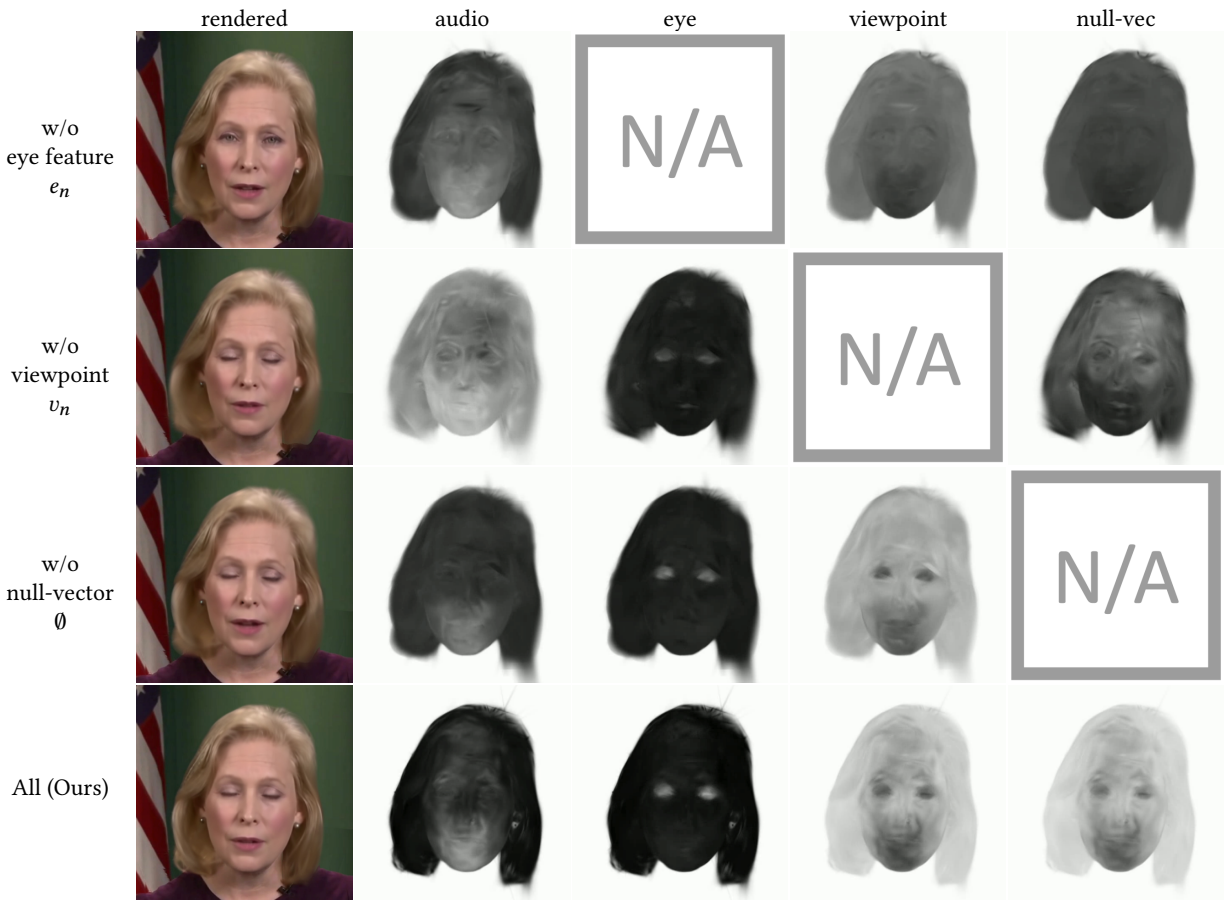


Figure A7: Ablation study on disentanglement effect of each input conditions. We assess the effectiveness of each input condition by alternatively turning them on and off, and visualizing the attention scores of each condition.



## E SUPPLEMENTARY VIDEO

To comprehensively visualize the efficacy of our proposed method in the domain of talking facial video synthesis, we prepared a supplementary video. This video encompasses the results and analysis of our experiments presented in the main paper and the supplementary document. We showcase talking head videos generated under both the **self-driven** and **cross-driven** settings and compare them with previous NeRF-based works [2, 4, 6]. We also demonstrate the effectiveness of our **spatial-audio cross attention module** by showing how the attention scores of each condition evolve as the scene progresses. Lastly, the video includes a set of ablation studies that systematically examine the impact of each component of our proposed method.

## F FURTHER DISCUSSIONS

### F.1 Ethical Considerations

Our goal with GaussianTalker is to create realistic talking 3D heads for practical real-world applications like digital assistants and video production. However, its photorealism raises ethical concerns, as it's difficult to distinguish real from synthetic videos. This can be used to create deepfakes, which are manipulated videos that can be used to spread misinformation or damage someone's reputation. To address this, we propose several measures: 1) informing users about video authenticity, 2) sharing our results with deepfake detection communities to improve detection algorithms, and 3) advocating for digital watermarks in real videos to deter misuse. Finally, we believe responsible use requires clear regulations to govern deepfakes on social media, protecting users from potential manipulation.

### F.2 Limitations and future work

GaussianTalker shares a common limitation with previous NeRF-based talking head synthesis methods: per-identity training. This restricts the model's ability to generalize to new identities, making data preparation for audio and eye features time-consuming. Additionally, free-viewpoint rendering remains a challenge due to the lack of multi-view training data. While the deformation stage achieves high fidelity and generalizes well to out-of-domain audio, it struggles with extreme viewpoints. Our current approach uses limited canonical training for coarse structure, leading to inconsistencies when synthesizing from very different angles.

Future work will focus on overcoming these limitations. We aim to explore techniques for multi-identity training and efficient data pre-processing. Additionally, we will investigate methods for free-viewpoint rendering using techniques like multi-view data acquisition or neural rendering approaches.

## REFERENCES

- [1] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*.
- [2] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. 2021. AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5784–5794.
- [3] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)* 42, 4 (2023), 1–14.

- [4] Jiahe Li, Jiawei Zhang, Xiao Bai, Jun Zhou, and Lin Gu. 2023. Efficient Region-Aware Neural Radiance Fields for High-Fidelity Talking Portrait Synthesis. *arXiv preprint arXiv:2307.09323* (2023).
- [5] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2020. 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*. 484–492.
- [6] Jiaxiang Tang, Kaisiyuan Wang, Hang Zhou, Xiaokang Chen, Dongliang He, Tian-shu Hu, Jingtuo Liu, Gang Zeng, and Jingdong Wang. 2022. Real-time Neural Radiance Talking Portrait Synthesis via Audio-spatial Decomposition. *arXiv preprint arXiv:2211.12368* (2022).
- [7] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. 2021. Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4176–4186.