

# Supplementary Materials: Speaker-specific Talking Head Synthesis via 3D Gaussian Splatting

Anonymous Authors

## A DETAILED NETWORK STRUCTURE

We provide the detailed network structures of Speaker-specific Motion Translator and Inpainting Generator in Figure 1 and Figure 2, respectively. In Figure 1, the transformer encoder has 4 layers, each with a 512 model dimension, 1024 inner dimension and 4 attention heads. The transformer decoder has 1 layer with the same parameter configuration as the encoder. MLP contains 3 hidden layers, and the channels of hidden layers are 512.

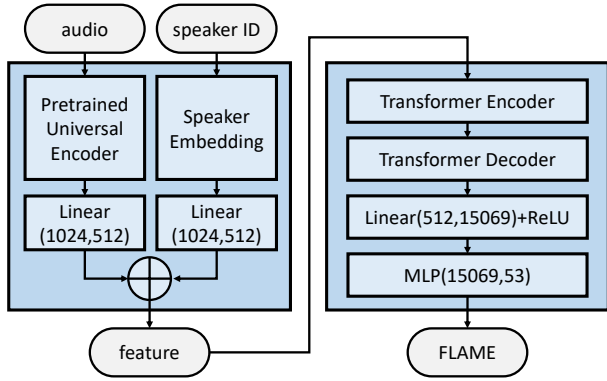


Figure 1: The detailed structure of Speaker-specific Motion Translator.

## B DETAILED EXPERIMENTAL SETTINGS

### B.1 Speaker-specific Motion Translator

During training of Speaker-specific Motion Translator, we set hyper-parameters of losses as follows:  $\lambda_v = 1000$ ,  $\lambda_y = 10$ ,  $\lambda_{sth} = 1000$ , and  $\lambda_{lat} = 0.001$ .

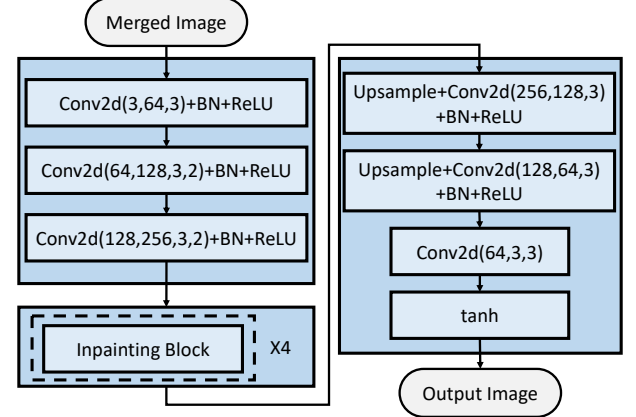
Table 1: Hyper-parameters of Dynamic Gaussian Renderer.

Name	$k$	$Q_{init}$	$\lambda_1$	$\lambda_2$	$\lambda_3$	$\lambda_p$	$\epsilon_p$	$\lambda_s$	$\epsilon_s$	$\lambda_{seg}$
Value	3	10522	1.0	0.1	0.2	0.1	0.05	0.1	2.0	0.5

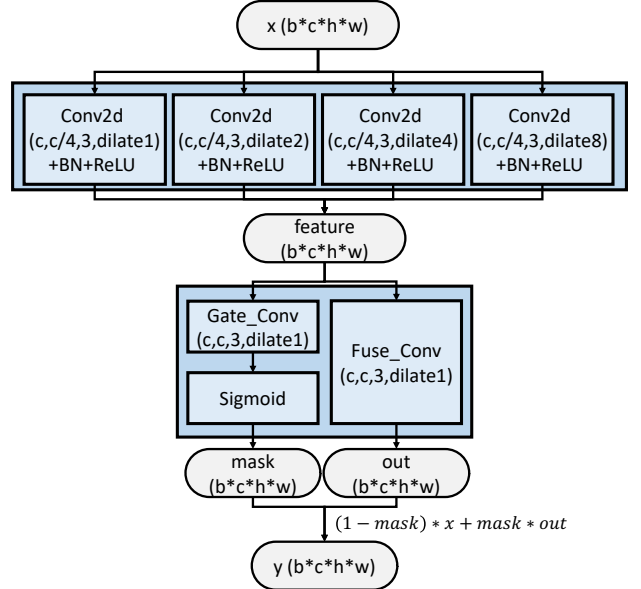
### B.2 Dynamic Gaussian Renderer

In Dynamic Gaussian Renderer, we set hyper-parameters in Table 1. Among these parameters,  $k$  denotes the order of the spherical harmonic function,  $Q_{init}$  denotes the number of initialized 3D Gaussians, and the remaining symbols are consistent with those described in the main text.

We also list the learnable parameters and networks in Table 2. Here,  $Q$  denotes the total number of 3D Gaussians, and  $T$  denotes the frame number of the video sequence. In the first part, we detail



(a) Inpainting Generator



(b) Inpainting Block

Figure 2: The detailed structure of Inpainting Generator. In subfigure (a) and subfigure (b), we show the overall structure of the model and the detailed structure of the Inpainting Block, respectively.

the attribute information of the 3D Gaussians. In the second part, we outline the network structure of the proposed Speaker-specific BlendShapes. In the third part, we present the designed BS weight information. In the fourth part, we enumerate the details of the FLAME parameters.

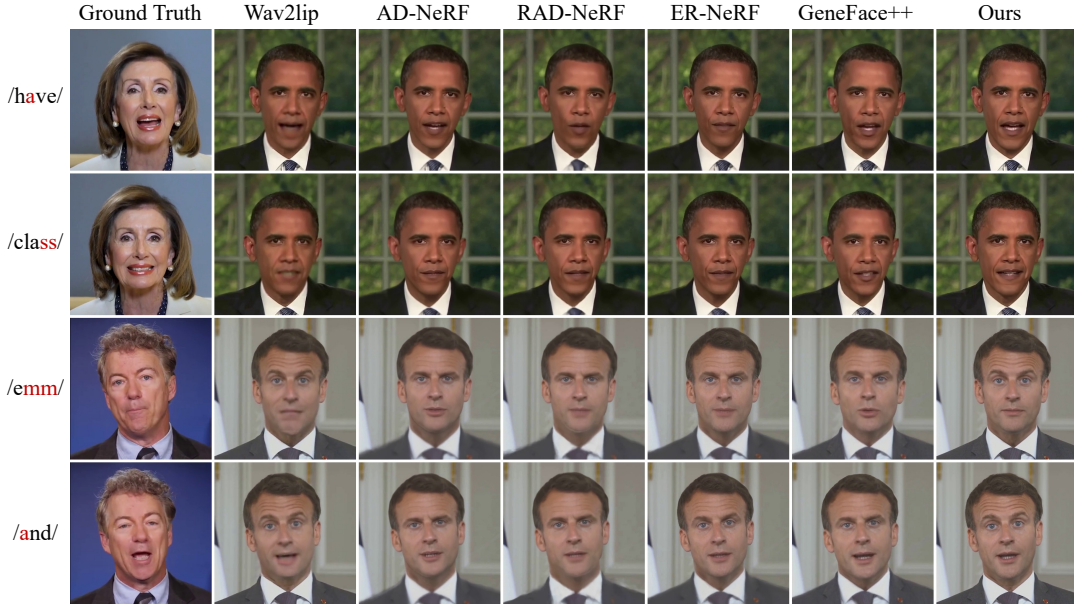


Figure 3: Comparison of six methods under the cross-driven setting. We show the results of head reconstruction and lip synchronization on Testset A and Testset B.

Table 2: Learnable parameters and Networks of Dynamic Gaussian Renderer.

Name	Shape	Learning Rate
$\tilde{u}$	$Q \times 3$	0.00016
$\tilde{s}$	$Q \times 3$	0.005
$\tilde{r}$	$Q \times 4$	0.001
$\alpha$	$Q \times 1$	0.05
$\kappa$	$Q \times 16$	0.0025
$W_\gamma$	Linear $16 \times 256$ ,	0.0001
	LeakyReLU,	
	Linear $256 \times 256$ ,	
	LeakyReLU,	
$W_{\gamma}$	Linear $256 \times 8$	0.0001
$W_{pos}$	$Q \times 3 \times 8$	0.00016
$W_{rot}$	$Q \times 4 \times 8$	0.001
$W_{color}$	$Q \times 1 \times 3 \times 8$	0.0005
shape	$T \times 100$	0.0001
expression	$T \times 50$	0.001
jawpose	$T \times 3$	0.0001
global orient	$T \times 3$	$1e-5$
translation	$T \times 3$	$1e-5$

## C ADDITIONAL EXPERIMENTS

### C.1 Additional Quantitative Evaluation

Our method only renders the speaker’s facial region, with the remaining regions utilizing the original reference frame. To more fairly evaluate the generated image quality, we re-crop all video

Table 3: Quantitative results under the self-driven setting with cropped face region. The best and second-best results are in bold and underlined.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	LMD $\downarrow$
Wav2Lip	28.8987	0.8911	0.1528	6.498
AD-NeRF	<u>29.6298</u>	<u>0.8933</u>	0.1104	4.620
RAD-NeRF	28.4605	0.8746	0.1010	4.070
ER-NeRF	28.7008	0.8820	<u>0.0608</u>	<u>3.668</u>
GeneFace++	25.6558	0.8263	0.1128	3.988
Ours	<b>33.2232</b>	<b>0.9504</b>	<b>0.0431</b>	<b>3.048</b>

frames to retain only the facial region. PSNR, SSIM, LPIPS, and LMD metrics for the cropped video frames are recalculated, as shown in Table 3. The results show that our method still has the best image quality considering only the facial region.

### C.2 Additional Qualitative Comparison

We show the comparison with other methods under the cross-driven setting in Figure 3. Regarding head reconstruction, our method significantly outperforms other methods in facial details, such as teeth. Concerning lip synchronization, our method demonstrates strong generalization capabilities with cross-identity and cross-gender audio. We strongly recommend watching our supplemental video for better visualization and more results.