Supplementary Material for TalkingGaussian

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

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

Overview

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

A Additional Experiments

A.1 Hybrid Motion Representation

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

A.2 Extension

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

⋆ Corresponding authors.

Table 5: Exploration of hybrid motion representations.

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

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

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

Audio Feature Extractor. Following previous baselines, we have adopted a pre-trained DeepSpeech [5] model to extract audio features in the main experiments, for a fair comparison. In fact, our method can also easily connect to more powerful feature extractors and become stronger. Table 6 shows our performance with different audio feature extractors Wav2Vec 2.0 [1] and HuBERT [6] under the *self-reconstruction setting*. The results demonstrate a high-quality audio feature could boost the performance of TalkingGaussian, especially on lip-synchronization, showing the growth potential of our approach.

Cross-Lingual and Cross-Gender. Our method can also applied to more challenging cross-lingual and cross-gender cases. In this experiment, we collect 4 training videos, in which 2 males and 2 females with English and French audio are included, and use challenging 3 test audio clips to drive their corresponding models. The three test audios consist of two German audio clips separately with a female and a male voice, and a Chinese audio clip with a male voice. In this setting, we compare our method to two baselines ER-NeRF [9] and GeneFace [14] that utilize different extractors.

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

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

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

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

B Implementation Details

B.1 Model Details.

Preprocessing. In the main paper, we use a frozen DeepSpeech [5] model to extract raw audio features. Then a CNN-based attention module in previous NeRF-based works [4,9,12,13] is adopted to adapt and smooth the audio features. The upper-face expression feature is composed of a set of action units that only relate to the upper-face motions, specifically: 1, 2, 4, 5, 6, 7, and 45. We use the action units detected by OpenFace [2, 3] as the input signals. Utilizing the same preprocessor in previous works [9, 13], we first estimate the head pose via a BFM [11] facial model, and then inversely calculate the camera pose.

Persistent Gaussian Fields. The Persistent Gaussian Fields are built on the 3DGS repository [6]. We init the fields with a random point cloud, and inherit the adaptive density control from 3DGS during training. For rendering, we adopt a modified 3DGS rasterizer [7], which enables alpha supervision, to penalize the empty areas.

Grid-based Motion Fields. In the experiments, we use three 2D hash-encoders [10] with a 3-layer MLP decoder to implement the Grid-based Motion Fields. For each 2D hash-encoder, we set the resolution range from 16 to 256 in the face branch, and 64 to 384 for the inside mouth branch. All these encoders are set with 12 levels. The hidden dimension of the MLP decoder is set to 64 and 32, respectively for the face and inside mouth branches.

Optimizer. During training, we keep two optimizers to optimize the Persistent Gaussian Fields and Grid-based Motion Fields separately. For the Persistent Gaussian Fields, we inherit the Adam optimizer from 3DGS with a similar hyperparameter setting to optimize the Gaussians. For the motion fields, we adopt an AdamW optimizer with learning rates of $5e{-}3$ and $5e{-}4$ for the hash encoder and other parts. An exponential scheduler is used to adjust the learning rates.

Metrics and Measurements. In the first setting, we measure the PSNR and LPIPS on the whole image, and SSIM on the face region. We also record the

[6] https://github.com/graphdeco-inria/gaussian-splatting
[7] https://github.com/ashawkey/diff-gaussian-rasterization

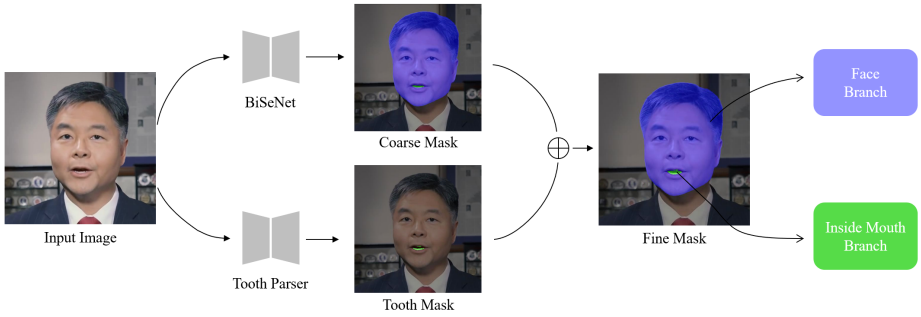

Fig. 8: Illustration of face and inside mouth segmentation.

person-specific training time and inference FPS of all methods. Since all the 2D-based generative baselines do not modify the upper part of the face, we do not measure their AUE-U. Notably, we provide another video clip as the image reference for Wav2Lip to avoid information leakage, for which PSNR, LPIPS, and SSIM are not valid. In the second setting, we use the non-comparison-based Sync-C and Sync-E to quantitatively measure the lip-synchronization quality. AUE-L is the sum of MSE on AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26. AUE-U is the sum of MSE on AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU45. The MSE loss of each AU item is averaged from all frames.

B.2 Face-Mouth Segmentation

As mentioned in Sec. 3.3, a semantic mask is used to divide the face and inside mouth in the 2D image. Specifically, we use a combination of two off-the-shelf face parsing models to generate the mask. First, we use a BiSeNet parser [15] pre-trained on CelebAMask-HQ dataset [8] to predict a coarse mask that contains a coarse mouth segmentation for the whole head and inside mouth. Due to the domain gap, the mouth mask from this face parser may fail to completely cover the mouth sometimes. To enhance the robustness, we further introduce a ResNet-based FPN trained on EasyPortrait [7] as a tooth parser to predict a tooth mask. Finally, we overlay these two masks to get a finer one, then use the corresponding masked image as $\mathcal{I}_{\text{mask}}$ in Eq. (10, 11) to supervise the two branches of TalkingGaussian. For each branch, we apply the semantic mask on both the ground truth image and the predicted result during training. The overall pipeline is shown in Fig 8.

C Additional Visualization

Here we show some high-definition synthesized frames in Fig. 9 for a convenient and intuitive visualization. Compared with current SOTA methods GeneFace [14] and ER-NeRF [9], our method performs best in image quality while

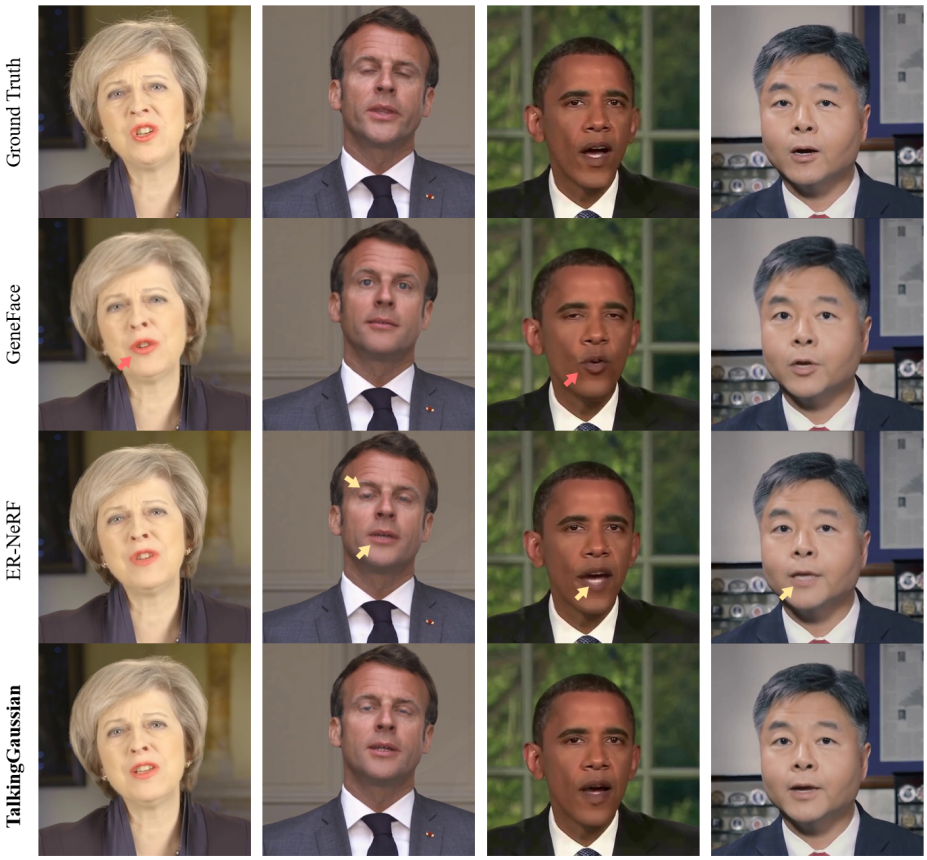

Fig. 9: Additional High-definition Comparisons. ER-NeRF [9] heavily sufferers the facial distortion problem caused by inaccurate appearance prediction. GeneFace [14] performs better in preserving fidelity, since it has introduced an intermediate representation to bridge the audio-visual mapping. However, its synchronization quality drops. In comparison, our method synthesizes better talking heads both in static and dynamic.

retaining a high lip-sync accuracy. We have also provided more additional results in our supplementary video. We strongly recommend watching it for better visualization.

D Dataset Declaration

In the experiments, all of the multimedia datasets we used were obtained from existing works [4, 9, 14]. To our knowledge, most of these data are collected from the internet. In our work, we have tried our best to use data containing only public figures to avoid invading personal privacy. All the data are manually checked to reduce the existence of offensive content.

E Ethical Consideration

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

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

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

F Limitations and Future Work

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

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

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

References

1. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020) 2
2. Baltrušaitis, T., Mahmoud, M., Robinson, P.: Cross-dataset learning and person-specific normalisation for automatic action unit detection. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). vol. 6, pp. 1–6. IEEE (2015) 3

3. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). pp. 59–66. IEEE (2018) 3
4. Guo, Y., Chen, K., Liang, S., Liu, Y.J., Bao, H., Zhang, J.: Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5784–5794 (2021) 3, 5
5. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al.: Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567 (2014) 2, 3
6. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3451–3460 (2021) 2
7. Kapitanov, A., Kvanchiani, K., Sofia, K.: Easyportrait - face parsing and portrait segmentation dataset. arXiv preprint arXiv:2304.13509 (2023) 4
8. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 4
9. Li, J., Zhang, J., Bai, X., Zhou, J., Gu, L.: Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7568–7578 (2023) 2, 3, 4, 5
10. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022) 3
11. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: 2009 sixth IEEE international conference on advanced video and signal based surveillance. pp. 296–301. Ieee (2009) 3
12. Shen, S., Li, W., Zhu, Z., Duan, Y., Zhou, J., Lu, J.: Learning dynamic facial radiance fields for few-shot talking head synthesis. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII. pp. 666–682. Springer (2022) 3
13. Tang, J., Wang, K., Zhou, H., Chen, X., He, D., Hu, T., Liu, J., Zeng, G., Wang, J.: Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. arXiv preprint arXiv:2211.12368 (2022) 3
14. Ye, Z., Jiang, Z., Ren, Y., Liu, J., He, J., Zhao, Z.: Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. In: The Eleventh International Conference on Learning Representations (2022) 2, 4, 5
15. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 325–341 (2018) 4