Monocular and Generalizable Gaussian Talking Head Animation

Supplementary Material

A. Limitations and Future Work

Despite the progress achieved in our talking head animation approach MGGTalk, several limitations warrant further investigation. We identify four primary areas for improvement: (1) Unnatural connection among the head, neck, and torso. Future work could employ a unified model of head, neck, and torso to enhance realism in their transitions and overall alignment. (2) Insufficient utilization of video information to improve context consistency. Incorporating multi-frame constraints during training could better estimate identity-specific shapes and maintain temporal coherence, thereby strengthening the naturalness of generated outputs [66]. (3) Potential inaccuracies in single-view depth estimation. Errors in depth estimates can compromise 3D modeling accuracy; adopting more robust approaches—such as DPT [67] or Sapiens [68]—may substantially improve reconstruction fidelity. (4) Unnatural results under severe asymmetry or challenging illumination. Complex lighting conditions can lead to unrealistic renderings, suggesting the need for illumination-aware control that adapts generation to diverse lighting environments. Addressing these limitations will further refine the quality, realism, and robustness of Talking Head Animation methods.

B. Ethics Consideration

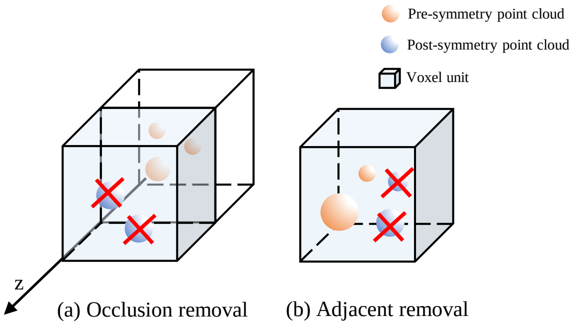
The proposed talking head animation method is primarily intended for applications in virtual communication and entertainment. Nonetheless, it may raise ethical and legal concerns if exploited for deceptive or harmful purposes by malicious actors. To mitigate these risks, it is essential to establish clear ethical guidelines and responsible usage practices that explicitly prohibit misuse. By doing so, we can help ensure that this technology is employed in a manner that promotes beneficial applications while minimizing potential harm.

C. Preliminary of 3DGS

3D Gaussian Splatting (3DGS) [21] utilizes anisotropic 3D Gaussians as geometric primitives to learn an explicit 3D representation. The geometry of each 3D Gaussian is defined as follows:

$$g(x) = e^{(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))} \tag{8}$$

where $\mu \in \mathbb{R}^3$ is the center of the Gaussian and $\Sigma \in \mathbb{R}^{3\times 3}$ is the covariance matrix that defines its shape and size. The covariance matrix Σ can be further decomposed into $\Sigma = RSS^TR^T$, where S denotes a scaling matrix determined by a scaling vector $s \in \mathbb{R}^3$, while R indicates a rotation matrix defined by a quaternion $r \in \mathbb{R}^4$. Additionally, each Gaussian has an opacity value $o \in \mathbb{R}$ which determines its visibility, and a color feature defined by $c \in \mathbb{R}^{12}$. Collectively, these parameters define each Gaussian as $\mathcal{G} = \{\mu, r, s, o, c\}$. Specifically, μ represents the position parameter of the Gaussian, which will be equivalently referred to as the three-dimensional coordinates of points in the Gaussian point cloud $\mathbf{P}$ in the subsequent discussion.



(a) Occlusion removal (b) Adjacent removal

Figure 9. Visualization of voxel filter rules. (a) illustrates occlusion removal along the z-axis, while (b) shows adjacent removal.

D. Implementation Details

D.1. Voxel-Based Filter

We propose a voxel-based filter $\mathcal{F}_{voxel}$ to effectively remove occluded or closely overlapping mirrored points from point clouds. The method voxelizes both the original and mirrored point clouds, then performs two key operations: z-axis occlusion detection and neighborhood occupancy checking. For z-axis occlusion, we compute the maximum z-values at each (x, y) voxel index in the original point cloud and compare them with mirrored points to discard occluded regions, as shown in Figure 9 (a). In the neighborhood check, we determine if mirrored points fall within the same voxel as the original points, considering them as neighboring points, and thus discarding them, as shown in Figure 9 (b). This combined approach efficiently retains essential mirrored points while removing those that are occluded or redundant.

D.2. Motion Deformation

The Motion Deformation module is designed to deform the 3D point cloud $[\mathbf{P}_f; \mathbf{P}_f^s]$ to synchronize it with the driving audio or driving image. Unlike directly editing the 3D point cloud using only the driving source, we also incorporate expression information from the source image to reduce the complexity of mapping arbitrary source expressions to arbitrary target expressions. We use 3DMM reconstruction [47] to extract the source expression basis β_s from the source image. For the driving input, the driving expression basis β_d is obtained via 3DMM reconstruction for driving images or an audio-to-expression method [7] for driving audio. The $\mathbf{MLP}$ then drives the source point cloud $[\mathbf{P}_f; \mathbf{P}_f^s]$ to generate the driven point cloud $[\mathbf{P}_d; \mathbf{P}_d^s]$ (Equation 9).

$$[\mathbf{P}_d; \mathbf{P}_d^s] = \mathbf{MLP}([\mathbf{P}_f; \mathbf{P}_f^s], \beta_d, \beta_s) \tag{9}$$

D.3. Gaussian Decoder

The Gaussian Decoder D_{gs} is designed to predict the remaining four Gaussian parameters—scaling $\mathbf{s}$, rotation $\mathbf{r}$, color $\mathbf{c}$, and opacity $\mathbf{o}$—for the visible deformed region point cloud $\mathbf{P}_d$. First, the input point cloud is reshaped into the form of a position map with dimensions (3, H, W). This map is then concatenated with the identity feature $\mathbf{F}$ extracted from the source image and fed into a UNet-based network to generate the $\mathbf{s}$, $\mathbf{r}$, $\mathbf{c}$, and $\mathbf{o}$. Finally, these maps are reshaped back into point cloud format and concatenated to form the complete set of Gaussian parameters $\mathcal{G}_d = \{\mathbf{P}_d \in \mathbb{R}^{H\cdot W\times 3}, \mathbf{s} \in \mathbb{R}^{H\cdot W\times 3}, \mathbf{r} \in \mathbb{R}^{H\cdot W\times 4}, \mathbf{c} \in \mathbb{R}^{H\cdot W\times 12}, \mathbf{o} \in \mathbb{R}^{H\cdot W\times 1}\}$.

D.4. Sym-Gaussian Decoder

The Sym-Gaussian Decoder D_{gs}^s is designed to generate Gaussian parameters for the non-visible regions of the point cloud $\mathbf{P}_d^s$. Given the challenge of obtaining sufficient information for these regions from the source image alone, facial symmetry priors are introduced as additional guidance. Specifically, the previously generated Gaussian parameters for the visible regions $\mathcal{G}_d$, identity features $\mathbf{F}$, and the symmetric point cloud $\mathbf{P}_d^s$ are concatenated and fed as input to a convolutional network to predict the offset relative to the already generated parameters. The networks for generating the biases of each Gaussian parameter are denoted as D_s^s, D_r^s, D_c^s, and D_o^s, respectively, and the generation process can be expressed as follows:

$$\begin{cases} \mathbf{s}^s = \mathbf{s} + D_s^s\left(\mathbf{F}, \mathbf{P}_d^s, \mathbf{s}\right) \\ \mathbf{r}^s = \mathbf{r} + D_r^s\left(\mathbf{F}, \mathbf{P}_d^s, \mathbf{r}\right) \\ \mathbf{c}^s = \mathbf{c} + D_c^s\left(\mathbf{F}, \mathbf{P}_d^s, \mathbf{c}\right) \\ \mathbf{o}^s = \mathbf{o} + D_o^s\left(\mathbf{F}, \mathbf{P}_d^s, \mathbf{o}\right) \end{cases} \tag{10}$$

Finally, we obtain the Gaussian parameters $\mathcal{G}_d^s = \{\mathbf{P}_d^s, \mathbf{s}^s, \mathbf{r}^s, \mathbf{c}^s, \mathbf{o}^s\}$ representing the non-visible facial regions of the source image.

D.5. Rendering and Inpainting

We use differentiable rasterization to render the Gaussian parameters $\mathcal{G}_{den}$ from the target viewpoint, resulting in an RGB image $\mathbf{I}_{tgt}^h$. To stabilize the training process, we additionally render the Gaussian parameters $\mathcal{G}$ before densification into a coarse image $\mathbf{I}_c^h$. Subsequently, we inpaint the torso and background regions of $\mathbf{I}_{tgt}^h$ using $\mathbf{I}_s^{bg}$, producing the final predicted image $\mathbf{I}_{tgt}$. Inspired by S^{3}D-NeRF [40], we employ a GAN-based network that takes the head image and the torso-background image as inputs to generate a 512 × 512 composite image.

E. Additional Results

To demonstrate the effectiveness of our approach, we provide additional visualizations and experimental results within the context of the video-driven talking head generation task.

E.1. Visualization of Gaussian Point Cloud Construction in DSGR

We utilize visualizations in Figure 10 to observe the three stages of point cloud construction in DSGR moudle. Initially, we use depth maps combined with normal maps as input for the Surface Reconstruction module, forming the initial facial geometry. Subsequently, a refinement network adjusts the initial construction, and facial symmetry is introduced to supplement the missing geometric structure in the occluded areas of the face.

For the first stage, the combination of depth and normal maps is critical. As illustrated in Figure 11, although the geometry derived directly from the depth map exhibits a stronger sense of three-dimensionality, it is often inaccurate due to monocular depth estimation limitations. For example, the first row shows an exaggerated nose, and the second row an overly sharp chin. Additionally, the geometric continuity of point clouds obtained through depth map back-projection is often inadequate, which hinders network training convergence. To address these inaccuracies, we incorporate normal maps to enhance geometric details. Both depth and normal maps are then used as inputs to the BINI algorithm [50] for surface reconstruction, producing more continuous and accurate 3D facial point clouds, as shown in Figure 11 and 10 (a), where surface reconstruction achieves smoother geometric continuity without the aforementioned structural inaccuracies.

The refinement network, detailed in the next phase, further adjusts the geometry to correct any residual inaccuracies, as illustrated in the Figure 10 (b). Although the initial point cloud constructed from depth and normal maps provides a good foundation, it may appear flat and fail to accurately represent the 3D facial structure. The refinement module effectively addresses these issues.

Finally, the application of symmetry plays a crucial role in reconstructing occluded regions of the face, which are often left incomplete in the initial stages. As shown in Figure 10 (c), The symmetry approach fills these gaps, ensuring a

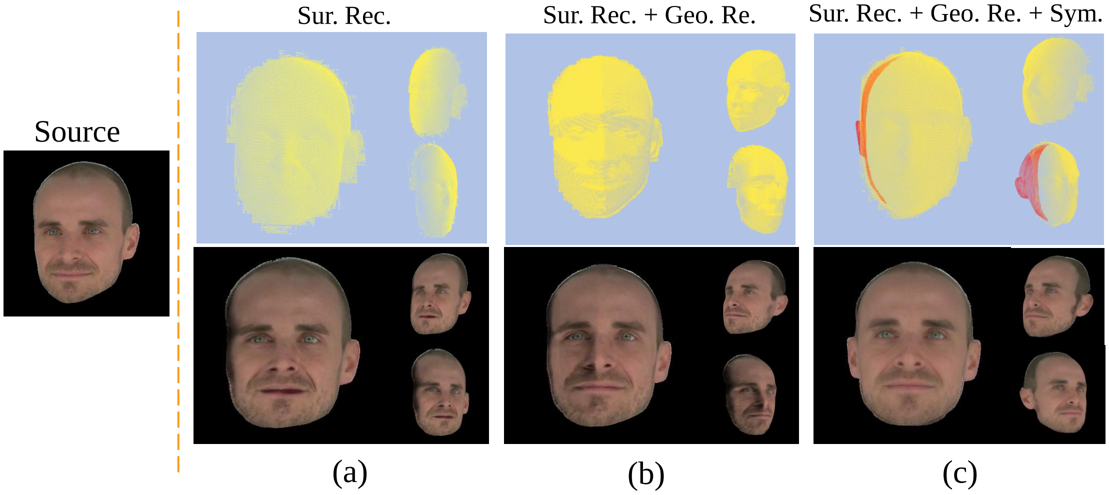


Figure 10. Visualization of the Gaussian point cloud construction in DSGR. The first row shows Gaussian point clouds obtained by ablating different construction modules, while the second row presents the corresponding rendered images from the SGP module. The yellow point cloud represents the geometry obtained from depth information, while the red point cloud indicates the symmetric augmentation of the geometry. Front, left, and right viewpoints are displayed.

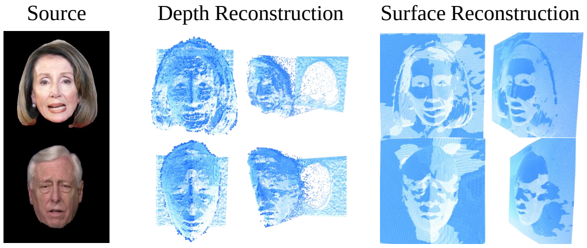


Figure 11. Visualization of geometric structures obtained from depth map projection versus surface reconstruction, including both frontal and side views of the 3D point cloud.

Table 5. Quantitative results of video-driven methods on the CelebV-HQ dataset [69]. We use **bold text** to indicate the best results and underline to denote the second-best results.

Methods	Self-Reenactment					Cross-Reenactment		
	PSNR↑	SSIM↑	FID↓	AED↓	APD↓	FID↓	AED↓	APD↓
Styleheat [10]	30.36	0.634	71.57	0.157	0.383	83.12	0.224	0.405
DaGAN [11]	30.81	0.626	57.72	0.113	0.196	60.45	0.244	0.308
ROME [53]	30.74	0.657	62.66	0.140	0.179	78.02	0.257	0.283
OTAvatar [54]	30.37	0.681	50.03	0.136	0.352	64.21	0.205	0.371
Real3DPortrait [18]	30.67	**0.696**	73.17	0.109	0.231	75.16	**0.191**	**0.254**
Portrait4D-v2 [14]	29.96	0.613	46.19	0.112	0.216	57.13	0.208	0.262
Ours	**30.84**	0.683	**42.23**	**0.104**	**0.173**	**56.43**	0.195	0.256

more comprehensive and accurate representation of the facial geometry across the entire point cloud.

E.2. Additional Results on HDTF and NeRSemble-Mono

In Figure 12, we present additional cross-identity reenactment results on the HDTF dataset (first four rows) and the NeRSemble-Mono dataset (rows five to eight). The results demonstrate that our method achieves strong identity consistency and 3D coherence, while effectively synchronizing facial expressions and poses with the driving source.

E.3. Additional Experiments on CelebV-HQ

Experimental Setups. To further evaluate the model's performance, we employed an additional dataset, CelebV-HQ [69], for quantitative and qualitative experiments on video-driven methods. Specifically, no training was conducted on this dataset; instead, 40 video clips were selected for inference. Data preprocessing and evaluation metrics were consistent with those used in the main text.

Quantitative Results. Experimental results on the CelebV-HQ dataset are presented in Table 5. For Self-Reenactment, our method outperforms others in appearance quality metrics (PSNR and FID) and is comparable to Real3DPortrait [18] in SSIM, indicating structural similarity. For Cross-Reenactment, our approach maintains a lead in FID, demonstrating superior identity preservation. Additionally, our AED and APD scores are close to Real3DPortrait [18], indicating effective control of facial expressions and poses.

Qualitative Results. The visual results on the CelebV-HQ dataset are shown in the last two rows of Figure 12. De-

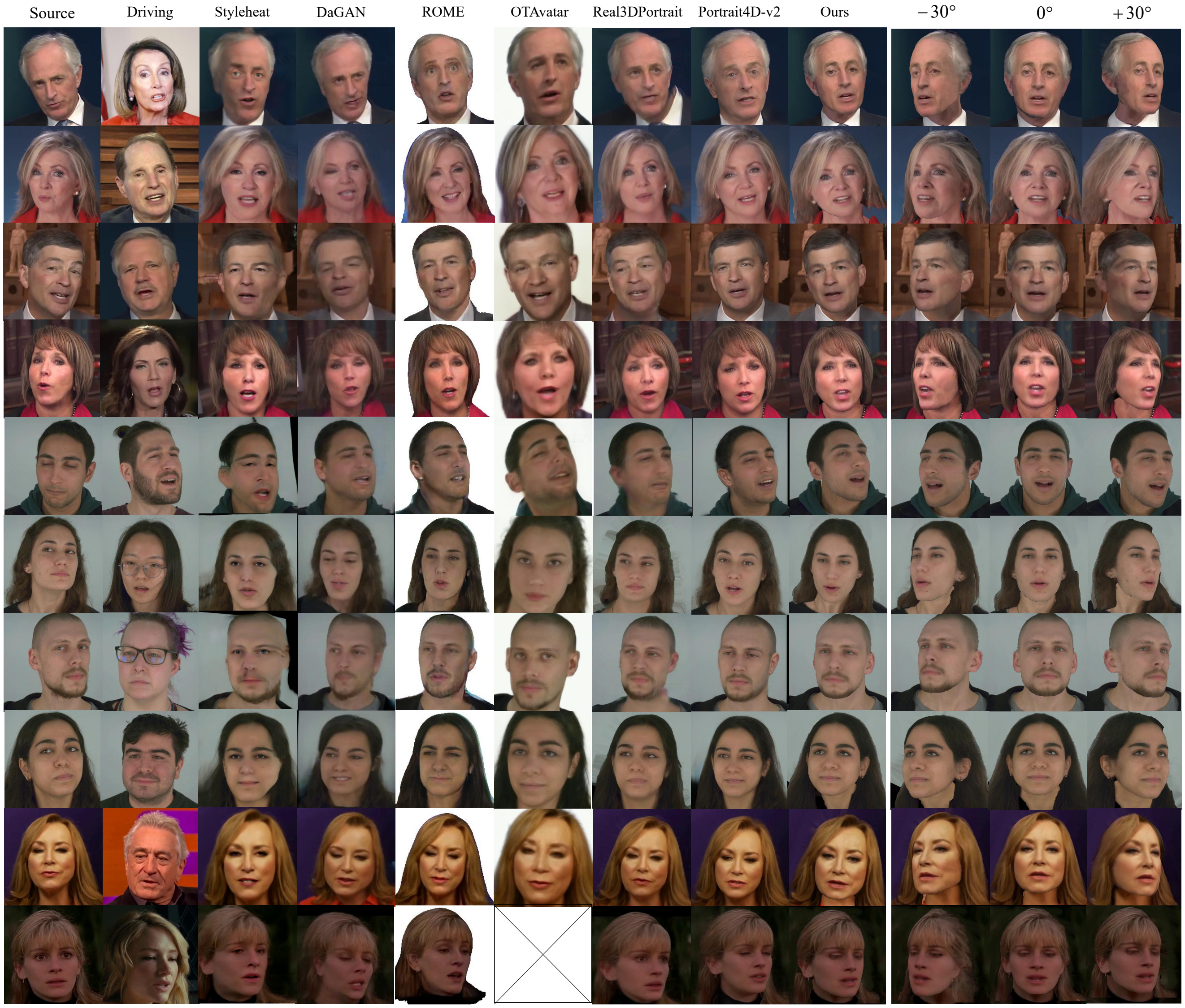


Figure 12. Qualitative comparisons with previous video-driven methods on the HDTF [51], NeRSemble-Mono [52] and CelebV-HQ [69] dataset. The first four rows show cross-identity driving results on the HDTF dataset, rows five to eight present results on the NeRSemble-Mono dataset, and the final two rows display results from the CelebV-HQ dataset. To demonstrate the multi-view consistency of our generated results, the last three columns display the fixed viewpoints at -30°, 0° and $+30^\circ$.

spite using significantly less training data compared to some methods [14, 18], our approach demonstrates competitive performance on unseen, in-the-wild dataset [69], maintaining strong 3D consistency as well as effective synchronization of expressions and poses.

E.4. Further Enhancement of Lip Synchronization

Our MGGTalk framework already achieves the second-best performance in terms of lip synchronization (LSE-C, LSE-D). As shown in Table 6, introducing SyncNet [64] provides additional performance improvements, suggesting that adopting an audio-based synchronization module can further refine the lip-sync accuracy of our method.

Table 6. Audio-riven results on HDTF [51] with the SyncNet [64] supervision.

Method	LSE-C↑	LSE-D↓
Wav2Lip [2]	8.84	6.48
MGGTalk	7.68	6.91
MGGTalk+SyncNet [64]	**8.87**	**6.35**

E.5. A Fairer Comparison with Lip-Sync Methods

In Table 2 of the main paper, we note that both Wav2Lip [2] and IP-LAP [27] rely on the ground-truth upper-half region to achieve pose alignment. To enable a more equitable comparison, we conducted experiments under a fixed pose setting, and as shown in Table 7, our method attains the highest image quality.

Table 7. Audio-driven results on HDTF [51] with fixed pose.

Method	PSNR↑	SSIM↑	FID↓	LMD↓
Wav2Lip [2]	30.02	0.664	30.53	3.94
IP-LAP [27]	29.47	0.631	36.14	3.87
Ours	**30.25**	**0.686**	**23.09**	**3.82**

E.6. Robustness of the Deformation Module

To evaluate the robustness of our Deformation module under inaccuracies in expression basis estimation, we introduce Gaussian noise into the expression basis and monitor the performance of the module. As shown in Figure 13 and Table 8, when the standard deviation of the Gaussian noise increases from 0 to 0.2, the predicted results remain relatively stable.

Table 8. Self-reenactment results on HDTF [51] with varying noise intensities added to the estimated expression features.

Noise std	FID↓	AED↓
0.00	18.95	0.102
0.05	19.13	0.104
0.10	19.46	0.117
0.20	19.74	0.121

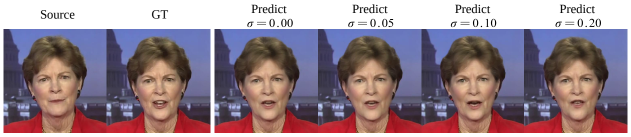


Figure 13. Visualization of adding noise to expression features.

References

[1] Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and CV Jawahar. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1428–1436, 2019. 1

[2] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM MM*, pages 484–492, 2020. 1, 5, 6, 8, 4

[3] Tianyi Xie, Liucheng Liao, Cheng Bi, Benlai Tang, Xiang Yin, Jianfei Yang, Mingjie Wang, Jiali Yao, Yang Zhang, and Zejun Ma. Towards realistic visual dubbing with heterogeneous sources. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1739–1747, 2021. 1

[4] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6):1–15, 2020. 1

[5] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *ECCV*, pages 35–51. Springer, 2020. 1

[6] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021. 1

[7] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661, 2023. 1, 2, 3, 5, 6

[8] Tao Liu, Feilong Chen, Shuai Fan, Chenpeng Du, Qi Chen, Xie Chen, and Kai Yu. Anitalker: Animate vivid and diverse talking faces through identity-decoupled facial motion encoding. *arXiv preprint arXiv:2405.03121*, 2024. 3, 6

[9] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024. 3, 6

[10] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European conference on computer vision*, pages 85–101. Springer, 2022. 2, 5, 6, 3

[11] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3397–3406, 2022. 2, 3, 5, 6

[12] Zhihua Xu, Tianshui Chen, Zhijing Yang, Chunmei Qing, Yukai Shi, and Liang Lin. Self-supervised emotion representation disentanglement for speech-preserving facial expression manipulation. In *ACM Multimedia 2024*, 2024.

[13] Tianshui Chen, Jianman Lin, Zhijing Yang, Chumei Qing, Yukai Shi, and Liang Lin. Contrastive decoupled representation learning and regularization for speech-preserving facial expression manipulation. *International Journal of Computer Vision*, pages 1–17, 2025. 1

[14] Yu Deng, Duomin Wang, Xiaohang Ren, Xingyu Chen, and Baoyuan Wang. Portrait4d: Learning one-shot 4d head avatar synthesis using synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7119–7130, 2024. 1, 3, 5, 6, 4

[15] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1606–1616, 2024. 2, 3

[16] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 2, 3

[17] Xiaozheng Zheng, Chao Wen, Zhaohu Li, Weiyi Zhang, Zhuo Su, Xu Chang, Yang Zhao, Zheng Lv, Xiaoyuan Zhang, Yongjie Zhang, et al. Headgap: Few-shot 3d head avatar via generalizable gaussian priors. *arXiv preprint arXiv:2408.06019*, 2024. 2, 3, 5

[18] Zhenhui Ye, Tianyun Zhong, Yi Ren, Jiaqi Yang, Weichuang Li, Jiawei Huang, Ziyue Jiang, Jinzheng He, Rongjie Huang, Jinglin Liu, et al. Real3d-portrait: One-shot realistic 3d talking portrait synthesis. *arXiv preprint arXiv:2401.08503*, 2024. 1, 3, 5, 6, 4

[19] Xueting Li, Shalini De Mello, Sifei Liu, Koki Nagano, Umar Iqbal, and Jan Kautz. Generalizable one-shot 3d neural head avatar. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3

[20] Tianshui Chen, Jianman Lin, Zhijing Yang, Chunmei Qing, and Liang Lin. Learning adaptive spatial coherent correlations for speech-preserving facial expres-

sion manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7267–7276, 2024. 1

[21] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 3, 1

[22] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2024. 2, 3

[23] Kartik Teotia, Hyeongwoo Kim, Pablo Garrido, Marc Habermann, Mohamed Elgharib, and Christian Theobalt. Gaussianheads: End-to-end learning of drivable gaussian head avatars from coarse-to-fine representations. *arXiv preprint arXiv:2409.11951*, 2024. 2

[24] Hongyun Yu, Zhan Qu, Qihang Yu, Jianchuan Chen, Zhonghua Jiang, Zhiwen Chen, Shengyu Zhang, Jimin Xu, Fei Wu, Chengfei Lv, et al. Gaussiantalker: Speaker-specific talking head synthesis via 3d gaussian splatting. *arXiv preprint arXiv:2404.14037*, 2024. 2, 3

[25] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2025. 2, 4, 5

[26] Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. One-shot talking face generation from single-speaker audio-visual correlation learning. In *AAAI*, volume 36, pages 2531–2539, 2022. 2

[27] Weizhi Zhong, Chaowei Fang, Yinqi Cai, Pengxu Wei, Gangming Zhao, Liang Lin, and Guanbin Li. Identity-preserving talking face generation with landmark and appearance priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2023. 3, 6, 5

[28] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13759–13768, 2021. 3

[29] Yifeng Ma, Suzhen Wang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, Zhidong Deng, and Xin Yu. Styletalk: One-shot talking head generation with controllable speaking styles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1896–1904, 2023. 2, 3

[30] Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1982–1991, 2023. 3

[31] Yifeng Ma, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yingya Zhang, and Zhidong Deng. Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv preprint arXiv:2312.09767*, 2023.

[32] Bohan Zeng, Xuhui Liu, Sicheng Gao, Boyu Liu, Hong Li, Jianzhuang Liu, and Baochang Zhang. Face animation with an attribute-guided diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 628–637, 2023. 3

[33] Antoni Bigata, Rodrigo Mira, Stella Bounareli, Konstantinos Vougioukas, Zoe Landgraf, Nikita Drobyshev, Maciej Zieba, Stavros Petridis, Maja Pantic, et al. Keyface: Expressive audio-driven facial animation for long sequences via keyframe interpolation. *arXiv preprint arXiv:2503.01715*, 2025.

[34] Gang Dai, Yifan Zhang, Qingfeng Wang, Qing Du, Zhuliang Yu, Zhuoman Liu, and Shuangping Huang. Disentangling writer and character styles for handwriting generation. In *CVPR*, pages 5977–5986, 2023.

[35] Gang Dai, Yifan Zhang, Quhui Ke, Qiangya Guo, and Shuangping Huang. One-dm: One-shot diffusion mimicker for handwritten text generation. In *ECCV*, pages 410–427, 2024. 3

[36] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009. 3

[37] Jiapeng Tang, Angela Dai, Yinyu Nie, Lev Markhasin, Justus Thies, and Matthias Nießner. Dphms: Diffusion parametric head models for depth-based tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1111–1122, 2024.

[38] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

[39] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Learning neural parametric head models. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 21003–21012, 2023. 3

[40] Dongze Li, Kang Zhao, Wei Wang, Yifeng Ma, Bo Peng, Yingya Zhang, and Jing Dong. Sˆ 3d-nerf: Single-shot speech-driven neural radiance field for high fidelity talking head synthesis. *arXiv preprint arXiv:2408.09347*, 2024. 3, 2

[41] Xuangeng Chu, Yu Li, Ailing Zeng, Tianyu Yang, Lijian Lin, Yunfei Liu, and Tatsuya Harada. Gpavatar: Generalizable and precise head avatar from image (s). *arXiv preprint arXiv:2401.10215*, 2024. 3

[42] Bo Chen, Shoukang Hu, Qi Chen, Chenpeng Du, Ran Yi, Yanmin Qian, and Xie Chen. Gstalker: Real-time audio-driven talking face generation via deformable gaussian splatting. *arXiv preprint arXiv:2404.19040*, 2024. 3

[43] Kyusun Cho, Joungbin Lee, Heeji Yoon, Yeobin Hong, Jaehoon Ko, Sangjun Ahn, and Seungryong Kim. Gaussiantalker: Real-time high-fidelity talking head synthesis with audio-driven 3d gaussian splatting. *arXiv preprint arXiv:2404.16012*, 2024.

[44] Yuelang Xu, Lizhen Wang, Zerong Zheng, Zhaoqi Su, and Yebin Liu. 3d gaussian parametric head model. In *European Conference on Computer Vision*, pages 129–147. Springer, 2025. 3, 5

[45] Tobias Kirschstein, Simon Giebenhain, Jiapeng Tang, Markos Georgopoulos, and Matthias Nießner. Gghead: Fast and generalizable 3d gaussian heads. *arXiv preprint arXiv:2406.09377*, 2024. 3

[46] Jiapeng Tang, Davide Davoli, Tobias Kirschstein, Liam Schoneveld, , and Matthias Niessner. Gaf: Gaussian avatar reconstruction from monocular videos via multi-view diffusion. 2025. 3

[47] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Computer Vision and Pattern Recognition Workshops*, 2019. 3, 5, 2

[48] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19680–19690, 2024. 3

[49] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. *arXiv preprint arXiv:2403.14627*, 2024. 3

[50] Xu Cao, Hiroaki Santo, Boxin Shi, Fumio Okura, and Yasuyuki Matsushita. Bilateral normal integration. In *European Conference on Computer Vision*, pages 552–567. Springer, 2022. 4, 2

[51] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 5, 6, 7, 4

[52] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 5, 6, 4

[53] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In *European Conference on Computer Vision*, pages 345–362. Springer, 2022. 5, 6, 3

[54] Zhiyuan Ma, Xiangyu Zhu, Guo-Jun Qi, Zhen Lei, and Lei Zhang. Otavatar: One-shot talking face avatar with controllable tri-plane rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16910, 2023. 5, 6, 3

[55] Jianghao Shen and Tianfu Wu. A pixel is worth more than one 3d gaussians in single-view 3d reconstruction. *arXiv preprint arXiv:2405.20310*, 2024. 5

[56] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[57] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 5

[58] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *ICCV*, pages 5784–5794, 2021. 5

[59] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[60] Connor Z Lin, David B Lindell, Eric R Chan, and Gordon Wetzstein. 3d gan inversion for controllable portrait image animation. *arXiv preprint arXiv:2203.13441*, 2022. 5

[61] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *ECCV*, pages 520–535, 2018. 5

[62] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[63] Gihoon Kim, Kwanggyoon Seo, Sihun Cha, and Junyong Noh. Nerffacespeech: One-shot audio-diven 3d talking head synthesis via generative prior. *arXiv preprint arXiv:2405.05749*, 2024. 6, 7

[64] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *ACCV*, pages 251–263. Springer, 2017. 7, 4

[65] Zhihao Xu, Shengjie Gong, Jiapeng Tang, Lingyu Liang, Yining Huang, Haojie Li, and Shuangping Huang. Kmtalk: Speech-driven 3d facial animation with key motion embedding. In *European Conference on Computer Vision*, pages 236–253. Springer, 2024. 8

[66] Ayush Tewari, Hans-Peter Seidel, Mohamed Elgharib, Christian Theobalt, et al. Learning complete 3d morphable face models from images and videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3361–3371, 2021. 1

[67] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 1

[68] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2024. 1

[69] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pages 650–667. Springer, 2022. 3, 4