Monocular and Generalizable Gaussian Talking Head Animation

Shengjie Gong[1]   Haojie Li[1]   Jiapeng Tang[2]   Dongming Hu[1]   Shuangping Huang[1,3†]
Hao Chen[1]   Tianshui Chen[4]   Zhuoman Liu[5]

[1]South China University of Technology   [2]Technical University of Munich   [3]Pazhou Laboratory
[4]Guangdong University of Technology   [5]The Hong Kong Polytechnic University
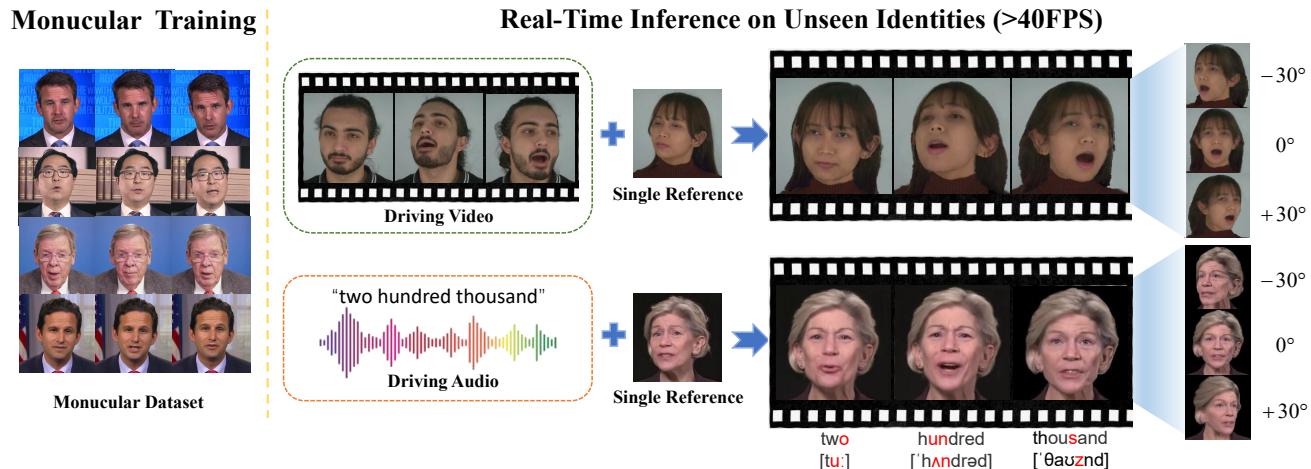
Figure 1. MGGTalk is trained using only monocular datasets and enable generalizing to unseen identities without personalized re-training. Additionally, it supports real-time generation of talking heads in diverse poses and novel viewpoints.

Abstract

In this work, we introduce Monocular and Generalizable Gaussian Talking Head Animation (MGGTalk), which requires monocular datasets and generalizes to unseen identities without personalized re-training. Compared with previous 3D Gaussian Splatting (3DGS) methods that requires elusive multi-view datasets or tedious personalized learning/inference, MGGtalk enables more practical and broader applications. However, in the absence of multiview and personalized training data, the incompleteness of geometric and appearance information poses a significant challenge. To address these challenges, MGGTalk explores depth information to enhance geometric and facial symmetry characteristics to supplement both geometric and appearance features. Initially, based on the pixelwise geometric information obtained from depth estimation, we incorporate symmetry operations and point cloud filtering techniques to ensure a complete and precise position parameter for 3DGS. Subsequently, we adopt a twostage strategy with symmetric priors for predicting the remaining 3DGS parameters. We begin by predicting Gaussian parameters for the visible facial regions of the source image. These parameters are subsequently utilized to improve the prediction of Gaussian parameters for the non-visible regions. Extensive experiments demonstrate that MGGTalk surpasses previous state-of-the-art methods, achieving superior performance across various metrics. Project page: https://scut-mmpr.github.io/MGGTalk-Homepage/.

1. Introduction

One-shot talking head generation synthesizes realistic talking head videos from a single reference image and driving sources like audio or motion. This technology is promising for applications such as video dubbing [1–3], film production [4], and video conferencing [5, 6]. A key challenge is generating high-quality, motion-rich videos from a single image, driven by expressive and pose-related information while ensuring robust identity protection.

Existing methods for talking head animation can be broadly categorized into 2D generator [2, 7–13] and 3D rendering [14–20] approaches. 2D generator methods typically use generative network, they often lack 3D modeling of the face and attribute disentanglement, which affects the overall generation quality. Some 3D rendering methods [14, 18, 19] utilize Neural Radiance Fields (NeRF) to synthesize more

---

†Corresponding author.

lifelike talking head videos by using 3D facial modeling. However, these NeRF-based approaches often suffer from high computational complexity, making real-time rendering challenging. In this context, 3D Gaussian Splatting (3DGS) [21] emerges as a promising solution, featuring a flexible, explicit, and high-fidelity 3D representation that allows for high-quality and fast rendering.

However, as shown in Figure 2, many current 3DGS-based methods [15, 16, 22–24] rely on multi-view datasets or personalized training to achieve high-quality results, which complicates generalization. Several methods [16, 23] generate realistic outputs but still depend on these multi-view datasets and personalized training approaches. To reduce data requirements, some approaches [15, 22, 24] utilize only monocular video data for personalized generation. Despite advancing data efficiency and accessibility, these methods still struggle to generalize to unseen individuals without additional training or fine-tuning. Another method [17] aims to balance data efficiency and generalization through few-shot learning techniques, generating satisfactory results with limited input. However, they still require multi-view data during training to build robust models, which constrains their effectiveness in real-world applications.

We propose the Monocular and Generalizable Gaussian Talking Head Animation (MGGTalk) framework, which is trained on monocular datasets and generalizes to new identities without additional fine-tuning. We exploit facial characteristics, focusing on depth estimation for capturing detailed facial geometry and leveraging prior knowledge of facial symmetry to improve the completeness of both geometric structure and texture. By integrating these characteristics, we address the incomplete feature representation often caused by monocular data limitations, generating more complete 3D Gaussian facial attributes similar to those achieved through multi-view or personalized learning approaches. In this way, we achieve a generalizable gaussian for high-quality, multi-view consistent talking head animation using only monocular data.

Specifically, we introduce two key modules: Depth-Aware Symmetric Geometry Reconstruction (DSGR), Symmetric Gaussian Prediction (SGP) . In the DSGR module, thanks to the advancements in monocular depth estimation, we use the pixel-wise depth estimation data obtained from Geowizard [25] as the initial information for 3D positions. Subsequent this initial capture is followed by a refinement network that mitigates common depth-related discrepancies, enhancing the precision of the resulting point cloud. To address the challenge of incomplete geometric information caused by extreme facial poses, we implement a symmetry operation in the canonical pose space to complement the geometry of regions that are invisible in the source image. Finally, we introduce a voxel-based filter that quan-
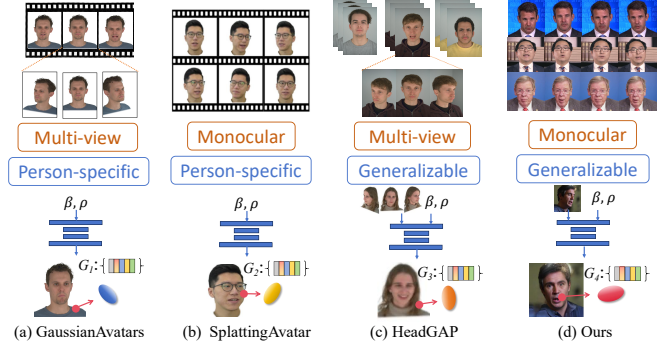

Figure 2. Comparison of 3DGS-based talking head animation methods, which typically rely on elusive multi-view datasets or tedious personalized learning. Our method achieves generalization to unseen identities while only training on monocular datasets.

tizes point cloud coordinates, effectively reducing redundancy and overlap in the reconstructed point cloud. Then, the SGP module is designed to better utilize geometric information and texture features to predict additional Gaussian parameters. Initially, we predict the remaining Gaussian parameters for the point cloud representing the visible facial regions of the source image, which benefits from labeled supervision, enabling more precise expression learning. Subsequently, for the non-visible areas derived from the symmetry operation in the DSGR model, we utilize the Gaussian parameters from the visible areas as inputs to guide the prediction process, thereby reducing the difficulty of predicting parameters for these unlabeled regions. This approach ensures that we obtain more complete and precise Gaussian parameters for the entire facial model.

Our main contributions are as follows:

- We propose the MGGTalk framework, which explores depth and symmetric facial characteristics for effective 3DGS-based talking head animation. It enables training on monocular datasets and generalizes to unseen identities without personalized re-training. To our knowledge, this is the first attempt to achieve this goal.
- We design the Depth-Aware Symmetric Geometry Reconstruction and Symmetric Gaussian Prediction modules, which leverage depth and symmetry information to enhance both geometric and appearance features for effective 3D Gaussian reconstruction and high-quality rendering.
- We conduct extensive experiments and ablative studies, demonstrating the effectiveness of MGGTalk in both audio-driven and video-driven talking head generation and exhibiting the actual contribution of each module.

2. Related Work

2.1. Talking Head Generation with 2D Generator

Many talking head generation methods utilize generative adversarial networks (GAN) [7, 10, 11, 26–29] or diffu-

sion models [8, 9, 30–35] as 2D generators to produce high-quality talking head videos. Some methods represent facial motion using keypoints [11, 27, 28], while others disentangle facial attributes in latent space for input to 2D generators [8]. However, the lack of 3D head modeling results in issues such as facial distortions and identity inconsistency. Although some approaches introduce 3D information, such as 3D Morphable Models (3DMMs) [36–39], to represent facial geometry or motion [7, 29, 32], the transformation of 3D features through 2D generators often results in a loss of geometric information, making it difficult to achieve consistent multi-view talking head generation. In contrast, our approach explicitly models the 3D face and utilizes point cloud-based rendering to generate multi-view consistent talking head images.

2.2. Talking Head Generation with 3D Rendering

Talking head generation methods that employ 3D rendering typically rely on neural or point-based techniques to achieve high-quality head avatars. Among these, Neural Radiance Fields (NeRF) have been widely adopted for generating detailed and realistic 3D facial reconstructions [14, 18, 19, 40, 41]. However, the computational complexity of NeRF leads to costly training and inference.

Recently, 3D Gaussian Splatting (3DGS) has demonstrated state-of-the-art visual quality and real-time rendering capabilities for 3D scene reconstruction tasks [21]. Many methods have successfully applied 3DGS for talking head generation, achieving impressive performance and rendering speed [15–17, 22, 24, 42–45]. Several of these approaches [15, 22, 24, 46] use several minutes of monocular video from a specific individual for training, achieving high visual quality and real-time rendering. For example, GAF [46] utilized multi-view head diffusion priors to reconstruct Gaussian avatars from monocular videos. SplattingAvatar [15] binds 3DGS to the 3DMM facial model to represent the head appearance and the facial expression changes are then realized by driving the positional offsets of 3DGS through the motion of the 3DMM. In contrast, GaussianAvatars [16] utilizes multi-view video to obtain a more accurate 3DMM and attaches 3DGS to the triangular mesh of the 3DMM, achieving improved 3D consistency. However, most of these methods require multi-view or monocular video from specific identities for training, limiting their ability to generalize to unseen individuals. Some methods have introduced 3D Gaussian parametric priors for pre-training, allowing generalization to new input images [17, 44]. Nevertheless, they require training on multi-view datasets, the acquisition of which is relatively demanding, thereby limiting their applicability. In this paper, we propose MGGTalk, which requires only monocular data for training and can generalize to unseen identities without additional retraining.

3. MGGTalk

Compared to multi-view or person-specific methods, monocular training with one-shot inference often leads to incomplete Gaussian parameters due to the limited information available. To address this challenge, we propose the MGGTalk framework, which uses Depth-Aware Symmetric Geometry Reconstruction (DSGR, Sec. 3.1) and Symmetric Gaussian Prediction (SGP, Sec. 3.2) modules to leverage facial depth and symmetry priors. This allows for the construction of a generalizable 3DGS representation from a single image, enabling high-quality, multi-view consistent one-shot talking head animation.

An overview of our framework is shown in Figure 3. Given a segmented head image from the source, the DSGR module uses depth information and symmetry priors to generate a point cloud representing the visible facial regions and a mirrored point cloud representing the invisible regions. Subsequently, an MLP-based deformation network takes expression features obtained from either the driving image or audio encoding as input and adjusts the point clouds, resulting in edited point clouds. Specifically, expression features are extracted using 3DMM estimation [47] for driving images or an audio-to-expression network [7] for driving audio. The SGP module then combines the identity encoding and the driven point cloud to generate the complete set of Gaussian parameters. Finally, the rendering and inpainting process projects the Gaussian parameters onto the target view and completes the torso-background, resulting in the final target image.

3.1. Depth-Aware Symmetric Geometry Reconstruction

Reconstructing 3D geometry from a 2D facial image is challenging due to the inherent lack of three-dimensional information, especially when parts of the face are invisible (e.g., in profile views). To address these challenges, we propose the Depth-Aware Symmetric Geometry Reconstruction module, which incorporates depth information and facial symmetry priors to obtain a more complete facial geometry from a single input image. Our approach reconstructs the 3D point clouds for the visible facial regions using depth estimation, while the symmetry operation supplements the geometry of the invisible regions.

For the visible facial regions of the input image, we introduce depth estimation information as a bridge between 2D and 3D. Some methods [48, 49] have already utilized depth maps to back-project into 3D space for constructing 3DGS positions. However, unlike the above methods that rely on multi-view depth estimation, monocular depth estimation often less accurate and lacks detailed information. To address this limitation, we employ a surface reconstruction module that takes depth and normal information derived from the depth estimation process to enhance surface
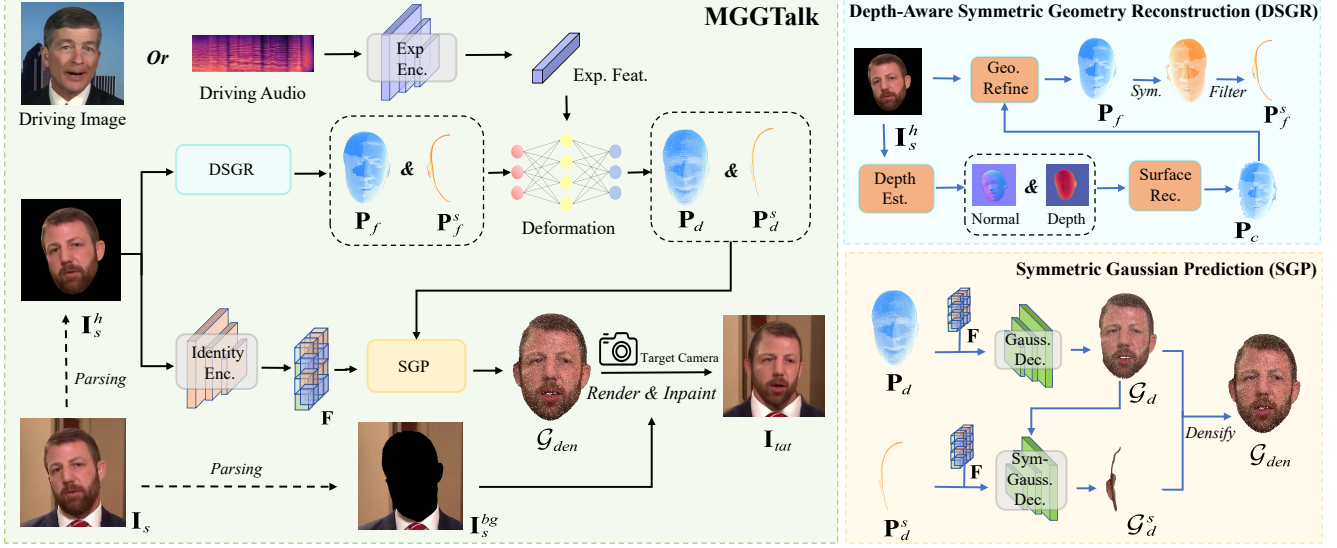
Figure 3. **Pipeline overview of the MGGTalk.** Given a source image $\mathbf{I}_s$, we first use semantic parsing to extract the head region $\mathbf{I}_s^h$ and torse-background $\mathbf{I}_s^{bg}$. The DGSR module generates point clouds $[\mathbf{P}_f; \mathbf{P}_f^s]$ for visible and invisible regions from $\mathbf{I}_s^h$. Expression features from the driving image or audio are used by the Deformation Network to edit the point cloud, resulting in $[\mathbf{P}_d; \mathbf{P}_d^s]$. The SGP module then takes the identity encoding $\mathbf{F}$ from $\mathbf{I}_s^h$ and the deformed point cloud $[\mathbf{P}_d; \mathbf{P}_d^s]$ to predict the complete Gaussian parameters $\mathcal{G}_{den}$. Finally, $\mathcal{G}_{den}$ is rendered and composited with torso-background $\mathbf{I}_s^{bg}$ to obtain the target $\mathbf{I}_{tat}$.

geometry and detail capture, providing a more robust geometric representation. A geometry refinement module is then introduced to achieve a more precise facial geometry.

Specifically, we first derive the depth map and normal map from the input image $\mathbf{I}_s^h$ using a pre-trained depth estimation Φ_{depth} [25]. We then apply the Bilateral Normal Integration algorithm (BINI) [50] on the depth map and normal map for surface reconstruction to generate the corresponding coarse facial point cloud $\mathbf{P}_c$. This process can be formalized as follows:

$$\mathbf{P}_c = \mathbf{BINI}(\Phi_{depth}(\mathbf{I}_s^h)). \qquad (1)$$

The initial geometric structure is relatively flat and fails to represent the 3D facial structure accurately. To address this, we introduce a 2D UNet-based geometry refinement network Φ_{refine}. This module takes the coarse geometry $\mathbf{P}_c$ and the source image $\mathbf{I}_s^h$ as inputs, predicting offset values for the 3D point cloud, which are then added to the coarse geometry to achieve refined corrections $\mathbf{P}_f$:

$$\mathbf{P}_f = \mathbf{P}_c + \Phi_{refine}(\mathbf{P}_c, \mathbf{I}_s^h). \qquad (2)$$

For the facial regions of the input image that are not visible, we utilize the natural left-right symmetry of the human face for supplementation. Specifically, we perform symmetry in the canonical pose space by inverting the x-coordinates of $\mathbf{P}_f$ to obtain a symmetry point cloud. We then propose a voxel filter $\mathcal{F}_{voxel}$ to remove overlapping regions between the symmetry point cloud and $\mathbf{P}_f$. The voxel filter quantifies the 3D point cloud coordinates. Points in the symmetry point cloud $\mathbf{P}_f^s$ are removed if they satisfy either

of the following conditions: (1) a corresponding point exists within the neighboring voxel of $\mathbf{P}_f$, or (2) they occlude some points in $\mathbf{P}_f$. The process of obtaining the symmetric point cloud $\mathbf{P}_f^s$ can be formulated as follows:

$$\mathbf{P}_f^s = \mathcal{F}_{voxel}(\mathbf{P}_f \cdot \left[-1, 1, 1\right]^T) \qquad (3)$$

3.2. Symmetric Gaussian Prediction

A straightforward approach to generate complete Gaussian parameters would be to concatenate the driven 3D point cloud $[\mathbf{P}_d; \mathbf{P}_d^s]$ with the identity features $\mathbf{F}$ and use a neural network to directly regress the Gaussian parameters. However, the symmetric portion of the point cloud represents regions that are not visible in the source image, making it difficult to acquire sufficient information to predict the corresponding Gaussian parameters. To address this challenge, we propose the Symmetric Gaussian Prediction module, which utilizes a two-stage approach for generating Gaussian parameters. First, the Gaussian parameters for visible regions are predicted. Subsequently, the facial symmetry prior is leveraged, along with the previously generated visible Gaussian parameters, to predict those for the symmetric, invisible regions. This approach mitigates the information deficiency for invisible regions and enhances the completeness of the Gaussian parameters.

Specifically, we first use the point cloud $\mathbf{P}_d$ and identity encoding $\mathbf{F}$ as inputs to the Gaussian Decoder D_{gs}, generating Gaussian parameters for the visible regions, denoted as $\mathcal{G}_d$. Next, considering the content similarity between the left and right facial regions, we use parameters from the visible regions as a foundation to infer the Gaussian parameters

5526

Table 1. Quantitative results of video-driven methods on the HDTF dataset [51] and NeRSemble-Mono dataset [52]. We use **bold text** to indicate the best results and underline to denote the second-best results.

Method	HDTF Self-Reenactment					HDTF Cross-Reenactment			NeRSemble-Mono Self-Reenactment					NeRSemble-Mono Cross-Reenactment		
	PSNR↑	SSIM↑	FID↓	AED↓	APD↓	FID↓	AED↓	APD↓	PSNR↑	SSIM↑	FID↓	AED↓	APD↓	FID↓	AED↓	APD↓
Styleheat [10]	30.23	0.664	61.92	0.209	0.245	76.43	0.369	0.334	30.67	0.719	77.45	0.189	0.602	93.25	0.345	0.854
DaGAN [11]	30.94	0.729	<u>33.23</u>	<u>0.106</u>	0.172	48.20	<u>0.225</u>	0.258	31.39	0.745	82.50	0.166	0.575	108.12	0.316	0.770
ROME [53]	30.99	0.775	76.44	0.135	<u>0.142</u>	79.31	0.229	0.286	31.07	0.712	95.27	0.142	0.420	119.09	0.288	0.551
OTAvatar [54]	30.67	0.654	36.47	0.133	0.226	50.37	0.210	0.312	31.23	0.695	59.42	0.158	0.331	73.30	0.304	0.492
Real3DPortrait [18]	<u>31.62</u>	0.712	33.26	0.145	0.163	51.36	0.256	0.293	<u>31.54</u>	0.667	79.09	<u>0.139</u>	<u>0.325</u>	88.91	<u>0.262</u>	<u>0.447</u>
Portrait4D-v2 [14]	30.12	**0.790**	36.57	0.111	0.158	<u>42.82</u>	0.238	<u>0.248</u>	30.29	**0.789**	<u>54.95</u>	0.182	0.367	<u>63.97</u>	0.275	0.522
Ours	**32.40**	<u>0.786</u>	**18.95**	**0.102**	**0.129**	**27.85**	**0.223**	**0.241**	**31.98**	<u>0.773</u>	**51.35**	**0.125**	**0.298**	**57.82**	**0.210**	**0.374**

for the invisible regions, which reduces the difficulty of directly predicting parameters for these unseen areas. Specifically we use the Sym-Gaussian Decoder D_{gs}^s, which takes the generated Gaussian parameters $\mathcal{G}_d$, the identity encoding $\mathbf{F}$ and the symmetric point cloud $\mathbf{P}_d^s$ as inputs to predict the symmetric Gaussian parameters $\mathcal{G}_d^s$ for the invisible regions. We concatenate the two sets of Gaussian parameters to obtain the complete Gaussian parameters $\mathcal{G} = [\mathcal{G}_d; \mathcal{G}_d^s]$. Notably, decoders D_{gs} and D_{gs}^s directly use the input dynamic point cloud as the Gaussian position output, while the scaling, rotation, color, and opacity parameters are generated by the network. Finally, to enhance the detailed rendering capabilities of the Gaussian representation, we apply a parent-child Gaussian densification process [55], where each Gaussian point in $\mathcal{G}$ generates a child node. The parent and child nodes together form the dense Gaussian representation $\mathcal{G}_{den}$. This process can be formulated as follows:

$$\mathcal{G}_d = D_{gs}(\mathbf{F}, \mathbf{P}_d) \tag{4}$$

$$\mathcal{G}_d^s = D_{gs}^s(\mathcal{G}_d, \mathbf{F}, \mathbf{P}_d^s) \tag{5}$$

$$\mathcal{G}_{den} = Densify([\mathcal{G}_d; \mathcal{G}_d^s]) \tag{6}$$

3.3. Loss Functions

During training, we synthesize both talking face image $\mathbf{I}_c^h$ and $\mathbf{I}_{tgt}^h$ which are rendered before and after densification. Several loss functions are utilized to ensure the similarity between synthesized image and ground truth image $\mathbf{I}_{GT}^h$, including L1 loss $\mathcal{L}_1$, SSIM loss $\mathcal{L}_{ssim}$ [56], and perceptual loss $\mathcal{L}_p$ [57]. The overall training objective is as follows:

$$\mathcal{L} = \mathcal{L}_1(\mathbf{I}_c^h, \mathbf{I}_{GT}^h) + \mathcal{L}_1(\mathbf{I}_{tgt}^h, \mathbf{I}_{GT}^h) +$$
$$\lambda_p[\mathcal{L}_p(\mathbf{I}_c^h, \mathbf{I}_{GT}^h) + \mathcal{L}_p(\mathbf{I}_{tgt}^h, \mathbf{I}_{GT}^h)] +$$
$$\lambda_{ssim}\mathcal{L}_{ssim}(\mathbf{I}_{tgt}^h, \mathbf{I}_{GT}^h) \tag{7}$$

with $\lambda_p = 0.01, \lambda_{ssim} = 0.2$.

4. Experiment

4.1. Experimental Setups

Datasets. We utilized the HDTF [51] and NeRSemble [52] datasets for training. After removing unstable footage, occlusions, and challenging lighting, approximately 400 video clips remained from HDTF. Unlike methods [17, 44] that use NeRSemble for multi-view supervision, we constructed monocular training pairs (source and driving images from the same video), resulting in around 300 video clips, denoted as NeRSemble-Mono. For preprocessing, we followed AD-NeRF [58] to crop video clips to 512×512 and estimate 3DMM pose parameters, then used Deep3DFaceReconstruction [47] to estimate 3DMM expression coefficients for each frame. For the HDTF and NeRSemble-Mono datasets, we use 80% for training and the remaining 20% for testing. To further assess the model's generalization ability, we collected in-the-wild talking videos for additional analysis.

Evaluation Metrics. We evaluate the methods in video-driven and audio-driven talking head generation, both using a single input image as the identity reference; the former employs a sequence of images, further divided into self-reenactment and cross-reenactment, while the latter relies on audio. To assess image quality, we employed metrics including PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index Measure) [56], and FID (Frechet Inception Distance) [59]. For the video-driven methods, we evaluate expression transition and pose reconstruction using AED (average expression distance)[60] and APD (average pose distance)[60], measured by a 3DMM estimator. For the audio-driven methods, we assess lip synchronization with LMD (Landmark Distance)[61] and audio-lip synchronization accuracy using LSE-C (Lip Sync Error Confidence)[2] and LSE-D (Lip Sync Error Distance)[2].

Implementation Details. During training, we freeze the Expression Encoder [7] and monocular depth estimation network [25]. The model is optimized using the Adam optimizer [62] with a batch size of 1 and a learning rate of

Figure 4. Qualitative comparisons with previous video-driven methods on the HDTF [51] and NeRSemble-Mono [52] dataset. The first two rows show the cross-identity driving results on the HDTF dataset, while the third and fourth rows present the results on the NeRSemble-Mono dataset. The last row shows the results of in-the-wild data. To demonstrate the multi-view consistency of our generated results, the last three columns display the fixed viewpoints at $-30°$, $0°$ and $+30°$.

$1e^{-4}$. The entire model, excluding the Inpainter, is trained for 1000K steps on a single RTX 4090 GPU, taking approximately 4 days, while the Inpainter is trained separately for 100K steps in under 8 hours. The final model achieves an inference speed exceeding 40 FPS on an RTX 4090.

4.2. Video Driven Talking Head Generation

Quantitative Results. We compare our method with state-of-the-art video-driven approaches, including 2D methods like Styleheat [10] and DaGAN [11], as well as 3D methods like ROME [53], OTAvatar [54], Real3DPortrait [18], and Portrait4D-v2 [14]. Results in Table 1 show our method achieves superior PSNR and FID scores in self-reenactment, confirming its effectiveness in high-quality texture generation and identity preservation. Our SSIM score ranks second, slightly behind Portrait4D-v2, highlighting our model's strength in reconstructing facial structure. In cross-reenactment, our approach maintains strong visual quality. Additionally, our method surpasses others on AED and APD metrics in both self- and cross-reenactment, demonstrating precise control of pose and expressions.

Qualitative Results. Figure 4 shows the qualitative results of cross-reenactment between different video-driven methods. These results show that we are able to effectively recover regions that are invisible in the source image while maintaining strong multi-view consistency. On one hand, the explicit decoupling of attributes in 3DGS allows for individual control over geometry, appearance, and pose. On the other hand, our in-depth exploration of facial information enables us to obtain relatively comprehensive Gaussian parameters even from a single reference image. For example, we observe that our method can generate the ear that is

not visible in the source image while maintaining high quality. Additionally, our method provides better control over facial pose and expression than the competing approaches. Lastly, we also present the results of our method on in-the-wild data. Despite using less than one-tenth of the training data compared to other methods [14, 18], our approach still demonstrates competitive performance.

Table 2. Quantitative comparisons with the state-of-the-art audio-driven methods on the HDTF dataset [51]. We use **bold text** to indicate the best results and underline to denote the second-best results. The methods [2, 27] highlighted against a gray background, utilize the ground truth upper half of the face as conditional input, focusing on the generation of the lower half of the face.

Method	PSNR ↑	SSIM ↑	FID ↓	LMD ↓	LSE-C ↑	LSE-D ↓
Ground Truth	N.A.	1.000	0.00	0.00	8.25	6.87
Wav2Lip [2]	34.31	0.937	18.05	3.39	8.84	6.48
IP-LAP [27]	31.67	0.874	26.62	3.30	7.07	7.89
SadTalker [7]	28.60	0.567	32.77	4.07	7.63	7.42
AniTalker [8]	30.42	0.620	25.74	3.87	6.12	8.23
EchoMimic [9]	29.65	0.655	22.33	3.79	6.26	8.77
Real3DPortrait [18]	30.66	0.689	39.10	**3.61**	6.57	8.01
NeRFFaceSpeech [63]	28.52	0.676	37.89	3.73	6.35	8.57
Ours	**30.92**	**0.731**	18.73	3.75	**7.68**	**6.91**

4.3. Audio Driven Talking Head Generation

Quantitative Results. We compare our method with existing state-of-the-art audio-driven approaches, including 2D image generator-based methods such as Wav2Lip [2], IP-LAP [27], SadTalker [7], AniTalker [8], and EchoMimic [9], as well as 3D rendering-based methods like Real3DPortrait [18] and NeRFFaceSpeech [63]. Quantita-

Figure 5. Qualitative comparisons with previous audio-driven methods on the HDTF [51] dataset. The last three columns display the fixed viewpoints at $-30°$, $0°$ and $+30°$.

Table 3. Ablation results on the HDTF dataset in self-reenactment setting. The first three rows show the ablation results of the Depth-Aware Symmetric Geometry Reconstruction (DSGR) module, while the fourth to sixth rows present the ablation results of the Symmetric Gaussian Prediction (SGP) module.

Module	Method	PSNR ↑	SSIM ↑	FID ↓	AED↓	APD↓
DSGR	w/o Geo. Refine.	29.98	0.705	24.82	0.157	0.232
	w/o Sym.	31.24	0.760	20.23	0.112	0.147
	w/o Gauss. Filter	29.56	0.724	27.42	0.148	0.180
SGP	w/o Pointclouds Input	30.78	0.776	22.05	0.124	0.134
	w/o Sym. Gauss. Dec.	32.12	0.761	19.74	0.116	0.151
	w/o $\mathcal{G}_d$ in Sym. Gauss. Dec.	31.68	0.764	19.36	0.109	0.142
	w/o $Densify$	30.32	0.778	21.58	0.138	0.133
	Ours	**32.40**	**0.786**	**18.95**	**0.102**	**0.129**

tive Evaluation on image quality and audio-lip synchronization is shown in Table 2, which indicates that our method reaches excellent performance across all image quality assessment metrics compared with leading one-shot audio-driven talking face generation methods. Regarding lip synchronization, although Wav2Lip performs better on the LSE-C and LSE-D metrics owing to its joint training with SyncNet [64], our method remains superior to other fully talking-head generation methods in these metrics. Furthermore, our method achieved competitive LMD scores, very close to the second-best method, NeRFFaceSpeech [63], indicating its ability to produce convincingly synchronized lip movements with the corresponding phonetic sounds.

Qualitative Results. To qualitatively evaluate different audio-driven methods, we provide sampled images of generated talking face videos in Figure 5. Specifically, the source image and the ground truth (GT) are shown in the first and second columns of the figure, followed by synthesized images from different methods. As observed, beyond the superior visual quality and identity preservation, our method achieves mouth shapes and poses closer to the GT, demonstrating enhanced pose control and lip synchronization capabilities.

4.4. Ablation Study

As discussed above, the use of symmetry is crucial for obtaining complete geometry and Gaussian parameters. Therefore, in the following section, we focus on the role of facial symmetry priors in the Depth-Aware Symmetric Ge-
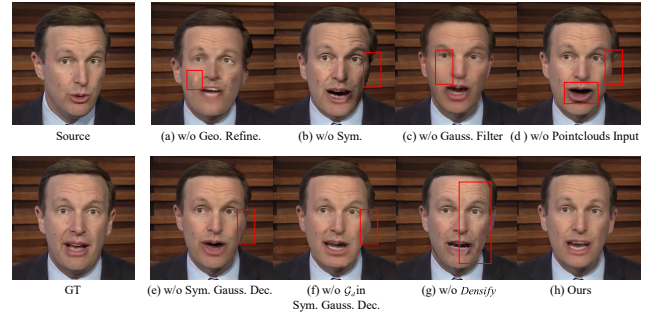

Figure 6. Visualization of ablation study results. The first image shows the source image and the ground truth. Figures (a) to (h) correspond to rows one through eight in the Table 3. The differences between images can be observed in greater detail by zooming in to 250%.

ometry Reconstruction (DSGR) and Symmetric Gaussian Prediction (SGP) modules, as well as ablate specific design choices within each module to further understand their impact on the overall performance. The quantitative results for the self-reenactment video-driven scenario are presented in Table 3, with the corresponding visual results shown in Figure 6.

4.4.1. Depth-Aware Symmetric Geometry Reconstruction

Effect of Symmetric Operation. To measure the importance of the symmetric operation, we removed it from the DSGR module (w/o Sym.). Since monocular depth estimation cannot provide depth information for the facial regions not visible in the source image, using only the point cloud obtained from depth estimation to represent the facial geometry results in incomplete modeling of the invisible regions. As shown in Figure 6 (b), the model fails to reconstruct the left ear that is occluded in the source image. To address this issue, we utilized symmetry to mirror the point cloud and fill in the missing regions. However, directly mirroring the point cloud can introduce significant overlap between the original and mirrored point clouds. Without filtering these overlapping regions (w/o Gauss. Filter), the two point clouds interfere with each other, leading to a loss of details in the generated images, as shown in Figure 6 (c). Table 3 also shows that without filtering to alleviate point cloud overlap, all metrics degrade significantly.

Effect of Refinement Network. When this refinement net-
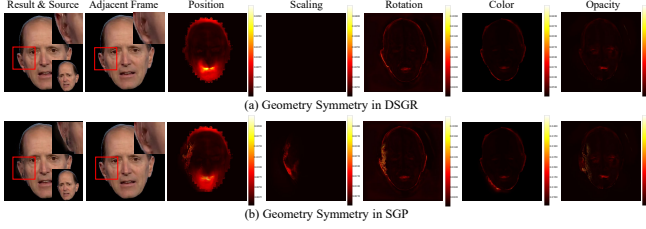
Figure 7. Impact of placing geometric symmetry in the DSGR or SGP module. The visualization shows two consecutive frames along with the variance of Gaussian parameters across all frames in the generated video. Differences between images can be examined in greater detail by zooming in to 350%.

Table 4. User study. We use **bold text** to indicate the best results and underline to denote the second-best results.

Method	Video-driven				Audio-driven			
	Ours	Portrait4D-v2	DaGAN	Others	Ours	EchoMimic	Wav2Lip	Others
ID. Pres.	**45.0%**	35.0%	5.0%	15.0%	**37.5%**	25.0%	17.5%	20.0%
Vis. Quality	**42.5%**	32.5%	5.0%	20.0%	**40.0%**	12.5%	25.0%	22.5%
Lip Sync.	**32.5%**	27.5%	12.5%	27.5%	35.0%	12.5%	**40.0%**	12.5%

work is removed (w/o Geo. Refine.), errors in the initial depth estimation propagate, affecting the prediction of all Gaussian parameters and resulting in degraded pose and appearance in the generated images, as shown in Figure 6 (a).

4.4.2. Symmetric Gaussian Prediction

Effect of Two-stage Generation Network for Symmetric Prior. We first replace the geometric information input in the SGP module with expression features directly (w/o Pointclouds Input). Without geometric awareness for each pixel, the model fails to reconstruct the invisible regions, and the visual quality in high-frequency motion areas, such as the mouth, also deteriorates (Figure 6 (d)). We further investigate the generation strategy for symmetric Gaussian parameters. As shown in Figure 6 (e)(f), when removing the Sym-Gaussian Decoder and using a single network to generate all Gaussian parameters simultaneously (w/o Sym. Gauss. Dec.), or using two separate networks to predict each part of the Gaussian parameters without providing the visible region's Gaussian parameters as input to the Sym-Gaussian Decoder (w/o $\mathcal{G}_d$ in Sym. Gauss. Dec.), they fail to generate well in the occluded regions of the source image. Due to the limited input information, directly learning parameters for occluded regions poses significant challenges. Using parameters from visible regions as guidance leverages their correlation, reducing the difficulty of prediction.

Effect of Densify Strategy. Removing the densification operation reduces generation quality, as demonstrated in Table 3 for quantitative results and visualized in Figure 6 (g), where the sparsity of points results in noticeable Moiré-like patterns in facial regions.

4.4.3. Impact of Geometry Symmetry Placement

This section explores the impact of applying geometric symmetry in the DSGR versus the SGP module. When symmetry is applied in the SGP module, the driven point cloud is influenced by driving expressions, leading to inconsistency in the mirrored geometry. This inconsistency causes significant variance in Gaussian parameters across frames, resulting in unstable visual output (Figure 7 (b)). To avoid this, we apply symmetry within the DSGR module to construct a consistent facial geometry before deformation. This approach ensures stable parameter prediction across frames and achieves a consistent facial appearance in the generated video (Figure 7 (a)).

4.5. User Study

We conducted a user study with 40 participants to evaluate the quality of talking head animations generated by different methods, using identical audio or video inputs for 30 clips. Participants were asked to evaluate three aspects: Identity Preservation, Visual Quality (intelligibility and naturalness), and Lip Synchronization. The questionnaire format is similar to KMTalk [65]. As summarized in Table 4, our method achieves the best results in both identity preservation and visual quality. For lip synchronization, our method ranks highest for video-driven inputs and second only to Wav2Lip [2] for audio-driven inputs.

4.6. Robustness of Asymmetric Heads

The SGP module generates Gaussian parameters for the invisible region based on the visible area guidance and symmetric point cloud, rather than simply copying parameters. Thus it can handle some head asymmetries (low reconstruction errors of asymmetric hairs in Fig. 8).
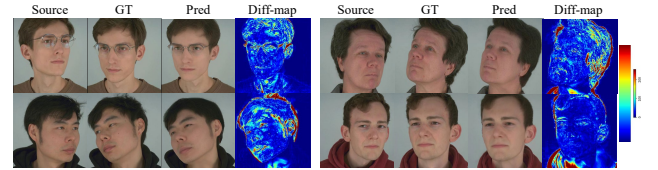

Figure 8. Qualitative results of asymmetric hairstyles. The Diff-map shows the squared difference between the GT and Pred.

5. Conclusion

We introduce MGGTalk, a novel framework that utilizes generalizable 3D Gaussian Splatting (3DGS) for one-shot talking head generation, trained exclusively on monocular data. The framework integrates facial depth geometry and symmetry priors across two core modules, enabling efficient prediction of 3DGS parameters and facilitating high-quality, multi-view consistent rendering. Evaluated on both audio-driven and video-driven tasks, MGGTalk demonstrates significant improvements in both quantitative and qualitative metrics compared to baseline methods.

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

[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 pretrained stylegan. In *European conference on computer vision*, pages 85–101. Springer, 2022. 2, 5, 6

[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

[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

[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 expression 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

[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

[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

[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

[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

[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

[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

[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

[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

[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 con-

trollable 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

[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