Wobble-Free 3D Talking Heads with Audio Driven Gaussian Splatting

Madhav Agarwal[1], Mingtian Zhang[2], Laura Sevilla-Lara[1], Steven McDonagh[1]
[1]University of Edinburgh, [2]University College London
{madhav.agarwal, l.sevilla, s.mcdonagh}@ed.ac.uk, mingtian.zhang.17@ucl.ac.uk

Abstract

Speech-driven talking heads have recently emerged, enabling the creation of realistic and interactive avatars. However, their real-world applications are limited, mainly because current methods are either very high-fidelity but slow, or fast yet temporally unstable. Towards addressing this issue, we create a robust and accurate lip-sync method, GaussianFlameTalk, that can generate person-specific avatars in real-time. Given a monocular video of a person and a potentially independent speech audio signal as input, our method generates temporally consistent talking head videos in real-time. Our pipeline invokes 3D Gaussian Splatting through a temporally consistent parameter mapping while also generating natural and believable lip-sync. By leveraging Gaussian Splatting mapped through 3D Morphable Models (3DMM), we introduce a novel transformer-based method, FlameTalk, to derive lip movement from audio. We directly predict the 3DMM parameters from audio, which are then used to control the rendering of a 3D Gaussian avatar. Further, we introduce a stability metric that can be used to quantify video output instability ("wobbling"). Our experimental work evidences state-of-the-art quantitative, qualitative performance for generative talking heads.

1. Introduction

Generating talking head videos, driven directly by audio, can be considered a highly valuable task with multiple practical applications [2, 5]. Whether in education, health care, teleconferencing, or the film and entertainment industry, high-quality personalized talking head avatars can serve as an effective path for information transfer. For instance, AI-driven virtual assistants for telemedicine can be useful in assistive communications and post-stroke rehabilitation [1]. The canonical problem involves taking an input video of a person, alongside an arbitrary audio speech signal, in order to create a person-specific avatar, capable of generating output video of the subject appearing to speak the audio content (*i.e.* with visual lip-syncing that matches the input audio).

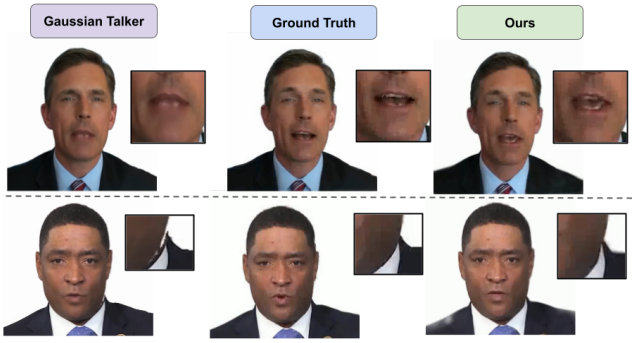

Figure 1. Our method can generate better lip movement, image quality, sharper teeth (1st Row), and reduced wobbling, artifacts (2nd Row) in comparison with GaussianTalker [7].

Previously proposed solutions for the problem involve using GANs [3, 21, 32, 37], Diffusion models [6, 38, 40], NeRFs [16, 20] and, more recently, 3D Gaussian Splatting [7, 8, 23, 29] based methods. However, there remain some shortcomings that prevent rendering high-quality video with photorealistic quality in real-time. The Diffusion-based methods generally possess state-of-the-art image quality, yet their inference time is slower than GANs. NeRFs provide faster rendering than Diffusion, yet they do not produce high quality image results, lacking details as well as real-time inference speed. 3D Gaussian Splatting (3DGS) has shown its efficacy in rendering high-quality images and videos. Taking advantage of this, there have been some recent advances in using 3DGS for creating audio-driven talking heads. Recent methods [7, 23] use a tri-planar representation to merge the audio signals directly into the learned 3DGS representation for rendering videos on a frame-by-frame level. However, the generated videos typically show flickering (or *wobbling*) in the facial region, causing visible artifacts. Our experiments show that this arises due to improper utilization of temporal information from an input video, which manifests as either inaccurate 3D parameter tracking of RGB videos or frame-by-frame generation, without any context of neighbouring frames.

To address this problem, we propose to process the audio signal using transformers [36] in a manner that can capture long-range semantic information [26, 33, 35]. We use

the input video to learn a person-specific style embedding, which can maintain the visual identity of the speaker. We note that directly mapping an audio signal to rasterised pixel space is difficult. We therefore alternatively opt to predict the FLAME [24] parameters for a template mesh and then use these to render the subject head via 3DGS [29]. We transfer the lip movement generated from a transformer-based network and head motion from the original video through an optimized set of FLAME parameters, obtained from training a GaussianAvatar. One aspect that is widely assessed when judging the quality of generated videos is that of stability [18, 31]. Intuitively; "*the video is stable*" is a fairly subjective statement. Can we formalize it? To answer this, we propose a stability metric for quantifying the temporal stability of avatar reenactment videos. Our contributions can be summarized as follows:

- We propose a novel component, FlameTalk, that uses a transformer-based architecture to generate FLAME [24] parameters directly based on input audio and a person-specific template mesh.
- We introduce a metric to quantify the temporal stability of synthetic talking head avatars.

We are the first to introduce a stable and complete talking head video generation architecture, GaussianFlameTalk, that predicts FLAME [24] parameters from audio and use 3D Gaussian Splatting to achieve real-time wobble-free video rendering, while maintaining photorealistic image quality, lip sync, and motion transfer through an optimized set of FLAME parameters in GaussianAvatar.

2. Related Work

2D Talking Head Generation Early 2D-image driven based talking head methods focus on utilizing a single image of a person and reenacting it with a driving video [3, 14, 19, 21, 32, 34, 37] using GANs [17]. Methods generally make use of an intermediate representation such as facial keypoints [3, 19, 21, 32, 37] or latent vectors [14, 34]. Some methods use audio to drive the motion, instead of a video [27, 42, 45]. Their focus is on achieving accurate lip-sync, while head-motion is generally hallucinated or learned from the training dataset. Given the superior generation quality of Diffusion methods [13], in comparison to GANs, some researchers recently employed them for face reenactment [6, 38, 40]. These methods provide better image quality, but the inference is typically slow and computationally expensive, making them infeasible for real-time generation.

3D Talking Head Generation With the advent of 3D rendering techniques such as NeRFs [25] and Gaussian Splatting [22], researchers have developed various methods to render talking heads. NeRF-based methods [16, 20] learn a radiance field from multiple input frames of a single scene. The volumetric rendering is performed based on the input controlling signal *e.g.* audio. Gaussian Splatting uses Gaussian optimization on input scene meshes. Rendering can be conditioned directly on audio or driving video [7, 8, 23, 29] for creating talking heads. Another line of work predicts only 3D Morphable Model (3DMM) parameters, such as FLAME [24], from an audio signal [12, 15, 30, 39]. We take advantage of an intermediate 3DMM representation to render Gaussian Splats in real-time.

3. Methodology

Our method is trained using an identity-specific video $V = \{I_n\}$ that consists of n frames. We train our model in a two-stage setting: the first stage (Sec. 3.1) involves training a person-specific Gaussian Splatting Model from the input video V, and the second stage (Sec. 3.2) involves learning an audio to FLAME [24] mapping that captures the speaking style of a given identity. The final video is generated by rendering the trained person specific Gaussian Avatar using the audio-driven learned FLAME parameters.

3.1. Gaussian Avatar Renderer

GaussianAvatar [29] introduced a method to explicitly bind Gaussians with the mesh triangles of FLAME [24] parametric model. The stability of the rendering process depends heavily on the accuracy of the binding. In previous work, such as INSTA [47], bounding volume hierarchy (BVH) [11] drives a nearest triangle search where warping leads to flickering artifacts. GaussianAvatar [29] is alternatively agnostic to the inaccuracy of input tracked meshes by allowing the back-propagation of positional gradients for each triangle. The consistent binding between Gaussians and mesh triangles, regardless of pose or expression, allows fine-tuning of FLAME parameters. Along with the optimization of Gaussian splats parameters for position and scaling, FLAME parameters (translation, rotation, and scaling) were also optimized during training. This plays a crucial role in stabilizing the output of rendering, as it mitigates misalignment between the meshes and Gaussians. We incorporated an identical Gaussian-based head modeling, where we replace the original tracked FLAME parameters with the optimized parameters obtained after training a person-specific avatar, to generate stable talking head avatars.

3.2. FlameTalk

We use FLAME [24], to map from an audio signal to facial motion. FLAME uses a set of disentangled parameters for controlling the identity, expression, and pose. These parameters are then used to generate a full 3D head mesh through trained GaussianAvatar. Distinct from previous work [15, 39], which operates directly on the full 3D head

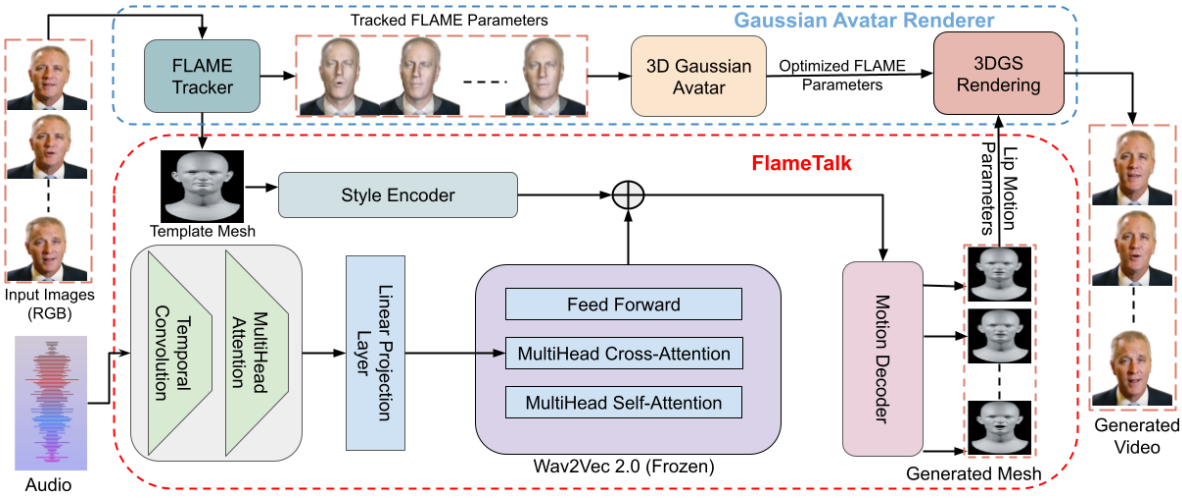

Figure 2. We introduce GaussianFlameTalk, which comprises of FlameTalk and Gaussian Avatar Renderer. We first generate meshes from an input video using VHAP [28] tracking. Given an input audio and a template mesh, FlameTalk uses a transformer-based architecture with a frozen Wav2Vec 2.0 [4] encoder. It learns long-term audio context and maps it directly to the 3D mesh by predicting FLAME [24] parameters. The generated parameters are used to render a person-specific GaussianAvatar [29], trained using the input video.

meshes by predicting triangle deformations or vertex positions, we take advantage of the disentangled FLAME representation to predict expression and pose parameters. By directly predicting FLAME expression parameters, we reduce the complexity of our learning objective from explicitly predicting the spatial location of thousands of face vertices to the prediction of fewer than one hundred parameters that together define facial expressions and lip motion.

We design a transformer-based architecture to capture long-range temporal information from the audio signal about the context of the spoken sentence. To mitigate the lack of diverse 3D audio-video datasets that contain 3D mesh information, audio and visual signals, we instantiate our encoder network using a pre-trained Wav2Vec 2.0 [4]. Wav2Vec 2.0 uses Temporal Convolution Layers, which encode audio signals into feature vectors. The model comprises of a stack of multi-head self-attention layers. We add a linear projection layer on top of the encoder to convert the output into a set of feature vectors. Similar to [15], we use a Periodic Positional Encoding (PPE) to provide temporal information to the decoder and a binary alignment mask to avoid information-leak from future frames.

For a single identity m, let the input training set be given by $L = \{A, M_{gt}^{1:T}, N_m\}$, where gt represents ground truth, $M_{gt}^{1:T}$ is a sequence of meshes for T frames and A is an audio signal from the ground-truth video corresponding to those frames. N_m represents a neutral template mesh for the given identity. Each input training set is generated by processing a video consisting of T frames using the VHAP tracker [29] to generate ground truth meshes $M_{gt}^{1:T}$ and neutral template mesh N_m. Our objective is to predict a sequence of meshes $M_{pred}^{1:T}$, given audio and neutral template mesh, such that: FlameTalk$(A, N_m) = M_{pred}^{1:T} \approx M_{gt}^{1:T}$.

The audio signal A is processed through the transformer encoder and a linear projection layer to generate audio features $C^{1:T}$. For a given frame t, the transformer encoder ingests audio for frames $\{1, \ldots, t\}$ and uses a linear projection layer to generate C^t. The predicted audio features are passed to the multi-head attention block of the transformer to obtain the latent vertex offsets $O_v^{1:T}$ for each frame. An identity-specific template mesh, which is an average of all the meshes obtained through video tracking, is encoded through a Style Encoder network, to obtain an identity embedding S. Predicted latent vertex offsets O_v^i for frame i are linearly combined with identity embedding as:

$$O_{sv}^i = S + O_v^i, \quad i \in \{1, \ldots, T\}. \tag{1}$$

These style-conditioned latent embeddings $O_{sv}^{1:T}$ are then processed by a motion decoder, which comprises a set of linear layers that map them to a low-dimensional FLAME parameter space, to obtain a 3D mesh representation. By performing this process for each frame i, we obtain a predicted mesh sequence $M_{pred}^{1:T}$.

Toward achieving accurate lip motion and jaw movement prediction, we isolate the parameters from the FLAME representation that are responsible for jaw movement. We use these to define an augmented ground-truth mesh, $M_{gt'}$ driven by FLAME parameter subset, and calculate a loss as the difference, in vertex space, between augmented ground-truth mesh and our predicted mesh per frame. The remaining FLAME parameter values, used to define the ground-truth mesh, are instantiated using the neutral template mesh, for every frame. The model is trained end-to-end using an $L2$ loss between the ground-truth and predicted meshes in

Paper	Self-Reenactment					Cross-Reenactment	
	PSNR↑	SSIM↑	LPIPS↓	Sync↑	Stability↓	Sync↑	Stability↓
IPLap [44]	29.0412	0.9462	0.0340	3.902	0.6633	3.324	0.6856
EDTalk [34]	26.9461	0.8626	0.0486	**7.144**	0.7802	**6.982**	0.7931
MimicTalk [41]	23.8775	0.8092	0.0735	5.446	0.8824	5.286	0.9227
GaussianTalker [7]	27.6079	0.9352	0.0451	5.346	1.7622	5.042	1.8745
TalkingGaussian [23]	27.3053	0.9335	0.0342	6.422	1.7183	6.146	1.8803
GaussianFlameTalk	**29.1233**	**0.9477**	**0.0338**	6.528	**0.6201**	6.122	**0.6836**

Table 1. Quantitative Comparison under Self-Reenactment and Cross-Reenactment. Our method, GaussianFlameTalk, achieves better results in terms of stability, realism and picture quality, and achieves comparable results for lip-sync with current state-of-the-art methods.

vertex space as follows

$$\mathcal{L}_{mesh} = \sum_{n=1}^{N} \left(\sum_{t=1}^{T} \left\| M_{gt'}^{t} - M_{pred}^{t} \right\|_2 \right). \quad (2)$$

During inference, FlameTalk ingests a neutral mesh and audio signal in order to predict a sequence of animated 3D facial meshes in FLAME parameters space. The predicted FLAME parameters are used to drive the motion of a person-specific avatar, which we instantiate in this work using GaussianAvatar [29], resulting in generating an audio driven talking head.

3.3. Quantifying temporal consistency

The considered existing works routinely generate talking head videos by posing video rendering as a set of, per-frame, independent tasks. We observe that this typically leads to poor temporal consistency in the output in the form of unnatural wobbling, aberrations, and face oscillations. Towards quantifying the problem, we adopt a strategy to measure the temporal smoothness of a given video.

We first select a video (ground-truth) and accompanying audio sample from the dataset [43]. We proceed to render a new talking head video using the original audio signal. Towards defining a robust evaluation protocol, we detect and track facial key points [46] on the nose, as these points are largely unaffected by jaw movement and expression changes. The time-domain signal, provided by these points, can then be compared between the generated and ground truth video frames. We observe that high-frequency wobbling and rapid oscillations are challenging to detect using keypoint comparisons alone, and adopt a hybrid approach by additionally performing a Fast-Fourier-Transform (FFT) analysis to identify frequent and uneven oscillations.

Our hybrid approach entails, for a given video, taking an average of the mean motion difference M_d, variability in motion magnitude V_m, and high-frequency power H_f. Each term is normalized by its respective maximum values across a given sequence of input frames. Our stability score is calculated by taking the average of these values, given by: Stability score $= (M_d + V_m + H_f)/3$

4. Experiments

Quantitative Evaluation We evaluate the performance of our model on two tasks: self-reenactment and cross-reenactment. In self-reenactment, we extract the last 30 seconds of the video and treat it as a test set. We train on the remaining video segment. For cross-reenactment, we use synthetically generated audio from a text-to-speech model[1] so that the audio sample has no identity information. We compare our method with state-of-the-art Gaussian Splatting [7, 23], GAN [34, 44] and NeRF [41] based methods. To evaluate self-reenactment, we use PSNR, SSIM, and LPIPS. We calculate the Sync confidence score [9, 10] and Stability score for both self-reenactment and cross-reenactment. Ou evaluation shows improvement over the state-of-the-art (Tab. 1). For perceptual metrics, GaussianFlameTalk performs better than NeRF and Gaussian-based methods. IPLap [44] gives comparable visual results, however, it is slower during inference. GaussianFlameTalk has lower, but comparable, lip-sync accuracy with EDTalk.

Qualitative Evaluation Fig. 1 shows visual comparisons of method performance. GaussianTalker and TalkingGaussian show poor lip sync quality, lower lip openness, and visual artifacts. They also show a lot of wobbling in the generated videos, mainly due to the lack of long-term temporal information and improper tracking of 3D parameters during training. Our method generates stable talking head videos, with qualitative imagery results that align with the relative quantitative metric improvements.

5. Conclusion

We introduce generate photo-realistic wobble-free 3D talking heads in real-time. Our pipeline improves temporal stability via a transformer that captures semantic information and long-range dependencies from the driving audio signal and introduce a stability metric to quantify the improvement. We report improved quantitative and qualitative performance over the state-of-the-art, showing our proposed method has potential utility in practical applications requiring both temporal stability and real-time inference.

[1] https://elevenlabs.io/

References

[1] Ayesha Afridi, Sumaiyah Obaid, Neha Raheel, and Farooq Azam Rathore. Integrating artificial intelligence in stroke rehabilitation: Current trends and future directions; a mini review. *JPMA. The Journal of the Pakistan Medical Association*, 75(2):445–447, 2025. 1

[2] Madhav Agarwal, Anchit Gupta, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. Compressing video calls using synthetic talking heads. In *British Machine Vision Conference (BMVC)*, 2022. 1

[3] Madhav Agarwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. Audio-visual face reenactment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5178–5187, 2023. 1, 2

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 3

[5] Bolin Chen, Jie Chen, Shiqi Wang, and Yan Ye. Generative face video coding techniques and standardization efforts: A review. In *2024 Data Compression Conference (DCC)*, pages 103–112, 2024. 1

[6] 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. 1, 2

[7] Kyusun Cho, Joungbin Lee, Heeji Yoon, Yeobin Hong, Jaehoon Ko, Sangjun Ahn, and Seungryong Kim. Gaussiantalker: Real-time talking head synthesis with 3d gaussian splatting. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 10985–10994, 2024. 1, 2, 4

[8] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1, 2

[9] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 251–263. Springer, 2017. 4

[10] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 87–103. Springer, 2017. 4

[11] James H Clark. Hierarchical geometric models for visible surface algorithms. *Communications of the ACM*, 19(10): 547–554, 1976. 2

[12] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael Black. Capture, learning, and synthesis of 3D speaking styles. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10101–10111, 2019. 2

[13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[14] Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8498–8507, 2024. 2

[15] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[16] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8649–8658, 2021. 1, 2

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[18] Wilko Guilluy, Laurent Oudre, and Azeddine Beghdadi. Video stabilization: Overview, challenges and perspectives. *Signal Processing: Image Communication*, 90:116015, 2021. 2

[19] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024. 2

[20] Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[21] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[22] 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

[23] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Talkinggaussian: Structure-persistent 3d talking head synthesis via gaussian splatting. In *European Conference on Computer Vision*, pages 127–145. Springer, 2024. 1, 2, 4

[24] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2, 3

[25] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[26] Ziqiao Peng, Yihao Luo, Yue Shi, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5292–5301, 2023. 1

[27] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, page 484–492, New York, NY, USA, 2020. Association for Computing Machinery. 2

[28] Shenhan Qian. Vhap: Versatile head alignment with adaptive appearance priors, 2024. 3

[29] 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. 1, 2, 3, 4

[30] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando de la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1173–1182, 2021. 2

[31] Marcos Roberto e Souza, Helena de Almeida Maia, and Helio Pedrini. Survey on digital video stabilization: concepts, methods, and challenges. *ACM Computing Surveys (CSUR)*, 55(3):1–37, 2022. 2

[32] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. 1, 2

[33] Wenfeng Song, Xuan Wang, Shi Zheng, Shuai Li, Aimin Hao, and Xia Hou. Talkingstyle: personalized speech-driven 3d facial animation with style preservation. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 1

[34] Shuai Tan, Bin Ji, Mengxiao Bi, and Ye Pan. Edtalk: Efficient disentanglement for emotional talking head synthesis. In *European Conference on Computer Vision*, pages 398–416. Springer, 2025. 2, 4

[35] Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20621–20631, 2023. 1

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[37] 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 Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2

[38] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024. 1, 2

[39] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12780–12790, 2023. 2

[40] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024. 1, 2

[41] Zhenhui Ye, Tianyun Zhong, Yi Ren, Ziyue Jiang, Jiawei Huang, Rongjie Huang, Jinglin Liu, Jinzheng He, Chen Zhang, Zehan Wang, et al. Mimictalk: Mimicking a personalized and expressive 3d talking face in minutes. *Advances in neural information processing systems*, 37:1829–1853, 2024. 4

[42] Yue Zhang, Minhao Liu, Zhaokang Chen, Bin Wu, Yubin Zeng, Chao Zhan, Yingjie He, Junxin Huang, and Wenjiang Zhou. Musetalk: Real-time high quality lip synchorization with latent space inpainting. *arxiv*, 2024. 2

[43] 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. 4

[44] 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 (CVPR)*, pages 9729–9738, 2023. 4

[45] 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. 2

[46] Zhenglin Zhou, Huaxia Li, Hong Liu, Nanyang Wang, Gang Yu, and Rongrong Ji. Star loss: Reducing semantic ambiguity in facial landmark detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15475–15484, 2023. 4

[47] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4574–4584, 2023. 2