

DGTalker: Disentangled Generative Latent Space Learning for Audio-Driven Gaussian Talking Heads

Xiaoxi Liang¹, Yanbo Fan^{2,*}, Qiya Yang¹, Xuan Wang³, Wei Gao¹, Ge Li^{1,*}

¹School of Electronic and Computer Engineering, Peking University,

²Nanjing University, ³Ant Group

1. Supplementary Material

In the supplementary material, we first present additional experimental results, including both quantitative and qualitative evaluations. We then provide further discussions on the audio feature extractor and the masked supervision strategy. Additionally, we include a dataset declaration and ethical considerations. Finally, we discuss the limitations of our work.

1.1. Experiment on Multi-view Dataset

To further validate rendering quality from novel viewpoints, we evaluated our method on the multi-view MEAD dataset [9]. In particular, we evaluate two identities by training the model solely on front-view clips and testing it on novel viewpoints, including top, down, left-30°, and right-30° viewpoints. The results are shown in Tab. 2. Notably, MEAD is not fully suitable for our experimental setting, primarily because: (i) the video clips are too short to support effective learning of audio generalization; and (ii) head rotation is entirely absent, making it difficult to capture accurate head geometry. As such, we present the results on MEAD merely for reference, in comparison with the latest state-of-the-art method, TalkingGaussian. Nevertheless, our method still demonstrates significant improvements under both frontal and novel views. We also advocate for the release of larger-scale and longer-duration multi-view talking head datasets to facilitate further research and development in this field.

1.2. Details of Quantitative Evaluation

In Tab. 1, we present results for four specific viewpoints and a spiral camera trajectory under the *novel-view self-reconstruction setting*. Our method consistently outperforms others across nearly all metrics and viewpoints. Although HFA-GP* utilizes the same Gaussian generative prior as ours, it significantly underperforms in both visual quality and synchronization metrics. Especially, its synchronization results lag substantially behind existing

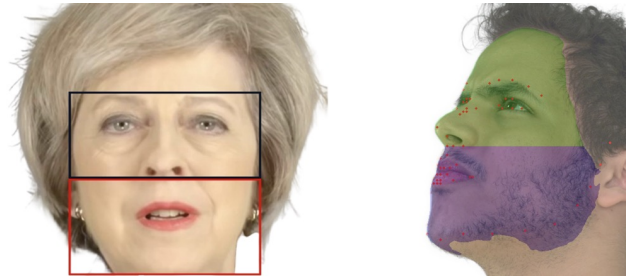


Figure 1. Two landmark region-based masked strategies.

monocular methods. These findings highlight that our superior performance stems from the effectiveness of the proposed approach.

1.3. Additional Qualitative Comparison

We further present an example of a spiral-view trajectory in Fig. 2, under the *novel-view self-reconstruction setting*. In this example, one frame is uniformly sampled every ten frames, arranged such that ground-truth images appear in odd-numbered rows and corresponding reconstructions in even-numbered rows. Our method demonstrates not only high-fidelity image quality across all viewpoints, but also preserves accurate motion dynamics and 3D consistency.

1.4. Supplementary Video

We present a supplementary video to comprehensively visualize the efficacy of our proposed method. This video includes: (i) video clips of two subjects in the *self-reconstruction setting*, showcasing both the ground truth view and a spiral camera trajectory view; (ii) a video clip in the *generalized 3D-aware audio-lip synchronization setting* with a spiral camera trajectory; and (iii) a video demonstrating the superior controllability of our method.

1.5. Audio Feature Extractor

Follow previous works [4, 5, 7, 10], we use the pretrained DeepSpeech [6] model to extract raw audio features for a fair comparison. We then process these features using the

*Corresponding authors

Methods	(+30°, +30°)			(+20°, +20°)			(-20°, -20°)			(-30°, -30°)			SpiralCamera		
	FID ↓	IDSIM ↑	Sync-C ↑	FID ↓	IDSIM ↑	Sync-C ↑	FID ↓	IDSIM ↑	Sync-C ↑	FID ↓	IDSIM ↑	Sync-C ↑	FID ↓	IDSIM ↑	Sync-C ↑
Ground Truth	N/A	1	8.468	N/A	1	8.468	N/A	1	8.468	N/A	1	8.468	N/A	1	8.468
ER-NeRF [7]	227.225	0.247	3.158	164.089	0.342	4.105	207.053	0.318	2.652	374.526	0.215	1.574	170.806	0.411	2.947
GaussianTalker [3]	191.177	0.277	3.343	111.590	0.310	4.408	120.267	0.355	3.833	182.188	0.304	3.384	86.440	<u>0.443</u>	3.900
TalkingGaussian [8]	167.810	0.291	5.609	124.132	0.363	5.890	101.666	<u>0.400</u>	5.296	216.066	0.334	<u>4.122</u>	79.897	0.428	<u>5.075</u>
HFA-GP* [1]	<u>135.908</u>	<u>0.348</u>	1.593	<u>77.867</u>	<u>0.388</u>	1.807	<u>90.453</u>	0.365	1.685	<u>126.064</u>	<u>0.337</u>	1.265	<u>67.711</u>	0.429	1.787
Ours	104.738	0.414	<u>5.354</u>	68.365	0.470	<u>5.411</u>	64.494	0.426	5.507	109.724	0.389	4.750	52.734	0.482	5.252

Table 1. Details of the quantitative comparison under the *novel-view self-reconstruction setting*. To save space, we report principle metrics over four specific viewpoints with yaw ($\pm 30^\circ$, $\pm 20^\circ$) and pitch ($\pm 30^\circ$, $\pm 20^\circ$), along with a spiral camera trajectory. The best and second-best methods are in **bold** and underline, respectively. Our results achieve the best performance across almost all metrics in almost all viewpoints.

Method	PSNR ↑	LPIPS ↓	Sync-C ↑	#.PSNR ↑	#.FID ↓	#.Sync-C ↑
Ground Truth	N/A	0	6.778	N/A	0	6.778
TalkingGaussian	23.365	0.134	2.659	18.823	115.189	1.957
Ours	26.475	0.089	2.903	21.574	66.347	2.111

Table 2. Results on the MEAD dataset. #. denotes the average performance across four viewpoints: top, down, left-30°, and right-30°.

same audio attention module, followed by a CNN-based module to adapt and smooth the audio features, as done in previous works [3, 8].

1.6. Masked strategy

Considering that speech activates the entire lower-face musculature rather than just the lips, we use the full lower-face region to supervise the lip vector in our masked cross-view supervision. Specifically, we adopt a simple strategy by extracting the upper-face and lower-face regions based on landmarks [2], as illustrated in the left part of Fig. 1. The black box denotes supervision on the upper face, while the red box indicates supervision on the lower face. This strategy works well across all tested videos. From a practical concern, especially when the video contains larger head rotations, we further integrate a face-parsing network with facial landmarks to achieve pixel-wise disentanglement, as illustrated in the right part of Fig. 1. In this figure, the green region denotes the upper-face area, while the purple region represents the lower-face area.

1.7. Dataset Declaration

In our experiments, all multimedia datasets were used from existing works [3, 5, 7, 10]. To the best of our knowledge, the majority of these datasets were collected from the internet. To address privacy concerns, we carefully curated datasets that predominantly feature public figures. Additionally, all data were manually reviewed to minimize the presence of inappropriate or offensive content.

1.8. Ethics Considerations

Our work is dedicated to developing highly realistic talking 3D heads for practical applications such as digital assistants and holographic communication. However, the photorealism achieved by our method introduces serious ethical concerns, as it becomes increasingly difficult to distinguish between real and synthetic content. This capability raises the risk of misuse—for example, the creation of deepfakes that could be employed to spread misinformation or damage individuals’ reputations. To address these concerns, we stress the importance of clearly communicating the synthetic nature of generated content to users. We also advocate for active collaboration with the deepfake detection research community to advance detection algorithms. As a preventive measure, we recommend embedding digital watermarks into authentic video content to deter misuse. Ultimately, we call for the establishment of clear regulations governing the use and dissemination of deepfake technologies on social media platforms to safeguard users from potential manipulation and exploitation.

1.9. Limitation and Future Work

We rely on a pretrained 3D GAN model to provide generative priors, which means that our rendering quality and generalization to novel views are inherently limited by the capacity of the underlying 3D GAN. Exploring end-to-end generation and driving frameworks remains a promising direction for future research.

In addition, our method focuses primarily on audio-driven precise facial motion, but does not capture hair and clothing dynamics, which are essential for realistic avatar applications. In future work, we plan to incorporate existing techniques for modeling hair and clothing dynamics to address this limitation.



Figure 2. Spiral-view qualitative results for the “Shaheen” identity under the *novel-view self-reconstruction* setting, with odd-numbered columns showing ground-truth images and even-numbered columns illustrating synthesized views along spiral camera trajectories.

References

- [1] Yunpeng Bai, Yanbo Fan, Xuan Wang, Yong Zhang, Jingxiang Sun, Chun Yuan, and Ying Shan. High-fidelity facial avatar reconstruction from monocular video with generative priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4541–4551, 2023. [2](#)
- [2] Lele Chen, Zhiheng Li, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII*, page 538–553, Berlin, Heidelberg, 2018. Springer-Verlag. [2](#)
- [3] 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, 2024. [2](#)
- [4] Lidong Guo, Xuefei Ning, Yonggan Fu, Tianchen Zhao, Zhuoliang Kang, Jincheng Yu, Yingyan Celine Lin, and Yu Wang. Rad-nerf: Ray-decoupled training of neural radiance field. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. [1](#)
- [5] 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](#)
- [6] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014. [1](#)
- [7] Jiahe Li, Jiawei Zhang, Xiao Bai, Jun Zhou, and Lin Gu. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7568–7578, 2023. [1](#), [2](#)
- [8] 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. [2](#)
- [9] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020. [1](#)
- [10] Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, Jinzheng He, and Zhou Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. *arXiv preprint arXiv:2301.13430*, 2023. [1](#), [2](#)