GGTalker: Talking Head Systhesis with Generalizable Gaussian Priors and Identity-Specific Adaptation

Supplementary Material

1. Additional Experiments

Additional Qualitative Experiments. To more intuitively compare the generalization to OOD speech, we present additional qualitative experiments in Fig. 1. The training data for the target identity contains only Chinese, but we use English during inference. As shown, previous methods often exhibit inaccurate lip shapes and blurred mouth regions, primarily due to the absence of Audio-Expression priors. Relying solely on a few minutes of speech features makes it difficult to generalize to different speaker voices and languages. In contrast, our method learns Audio-Expression priors from large-scale datasets and finetunes for the specific identity's style, enabling accurate lip synchronization even in cross-language scenarios.

Additional Ablations. During Customized Adaptation, our color MLP generate fine-grained, motion-aligned textures. Removing it would result in smoother textures, such as wrinkles (shown in Fig. 2).

2. Additional Implementation Details

2.1. Network Architectures

Identity-Gaussian Generator. In our Expression-Visual stage, we use an Attention-UNet as the Identity-Gaussian Generator. It consists of 5 downsampling layers, 4 upsampling layers and 4 attention gates, as shown in Fig. 3. It produces a feature map M of size 512×512, resulting in 262k Gaussians.

Body Inpainter. Our Body Inpainter blend the rendered results and the original background. It consists of 4 downsampling layers and 4 upsampling layers, as shown in Fig. 3.

2.2. Implementation Process

Audio-Expression Prior stage. We use HDTF [21], CN-CVS [2], and a self-collected 100-hour internet dataset, all of which are in-the-wild data. HDTF contains over 300 speakers with approximately 16 hours of English video. CN-CVS is a large Chinese audio-visual dataset with over 2,000 identities and more than 200 hours of data. We select 100 identities, resulting in about 27 hours of data. Additionally, we collect around 100 hours of multilingual data from the internet, primarily news broadcasts and speech videos, where audio and lip movements are highly synchronized. All data are cropped and resized to 512×512 with the head centered.

Expression-Visual Prior stage. Our expression includes 53 dimensions of FLAME parameters. Although the officially released "expression" from FLAME has only 50 dimensions [11], previous work has shown that the 3-dimensional jaw pose is highly correlated with expression and not disentangled [4]. Therefore, we jointly train "expression" and "jaw pose." To ensure 3D consistency, we use only 0th-order SH parameters to represent color, instead of the 3rd-order SH used in the original 3DGS [6]. We use the VFHQ [17] and NeRSemble [8] datasets. VFHQ is an in-the-wild dataset containing 15k identities, each with hundreds of video frames. NeRSemble is a high-resolution lab dataset recorded with 16 cameras, containing over 200 identities. We use both VFHQ and NeRSemble to expose the model to a large variety of identities while incorporating 3D priors. Following Qian et al., we use VHAP [14] to extract FLAME and camera parameters. VHAP is highly accurate for multi-view videos but can cause frame jitter and unnatural expressions in monocular videos. However, since our Expression-Visual Prior phase learns a coarse prior, both qualitative and quantitative results show that even with imperfect tracking, we can still learn sufficiently robust 3D head priors. During deployment, we use joint fine-tuning to mitigate tracking errors.

Order	Modules	Time
1	M_{id}	2 min
2	$M_{id} + F$	3 min
3	A2E	5 min
4	A2E $+ M_{id} + F$	5 min
5	A2E $+ M_{id} + F + \mathcal{M}_{\mathbf{SH}} + \mathcal{I}$	5 min
Total		20 min

Table 1. **Components optimized at each stage and the corresponding required time During customized adaption.** "A2E" refers to Audio-Expression Finetuning $\mathcal{F}$ denotes the FLAME parameters, containing FLAME shape, pose and expression .

Customized Adaptation stage. Our optimization steps are shown in the Tab 1. We sequentially optimize the Gaussian UV map, FLAME parameters, Audio-Expression, color MLP, and Body Inpainter. Any components not mentioned for each phase are frozen. In all experiments, we use the original video's head pose and eye movements, but we can also arbitrarily control head pose and blinking.

2.3. Implementation of Other Methods

For a fair comparison, we re-implement previous methods to conduct self-reenactment and cross-reenactment ex-

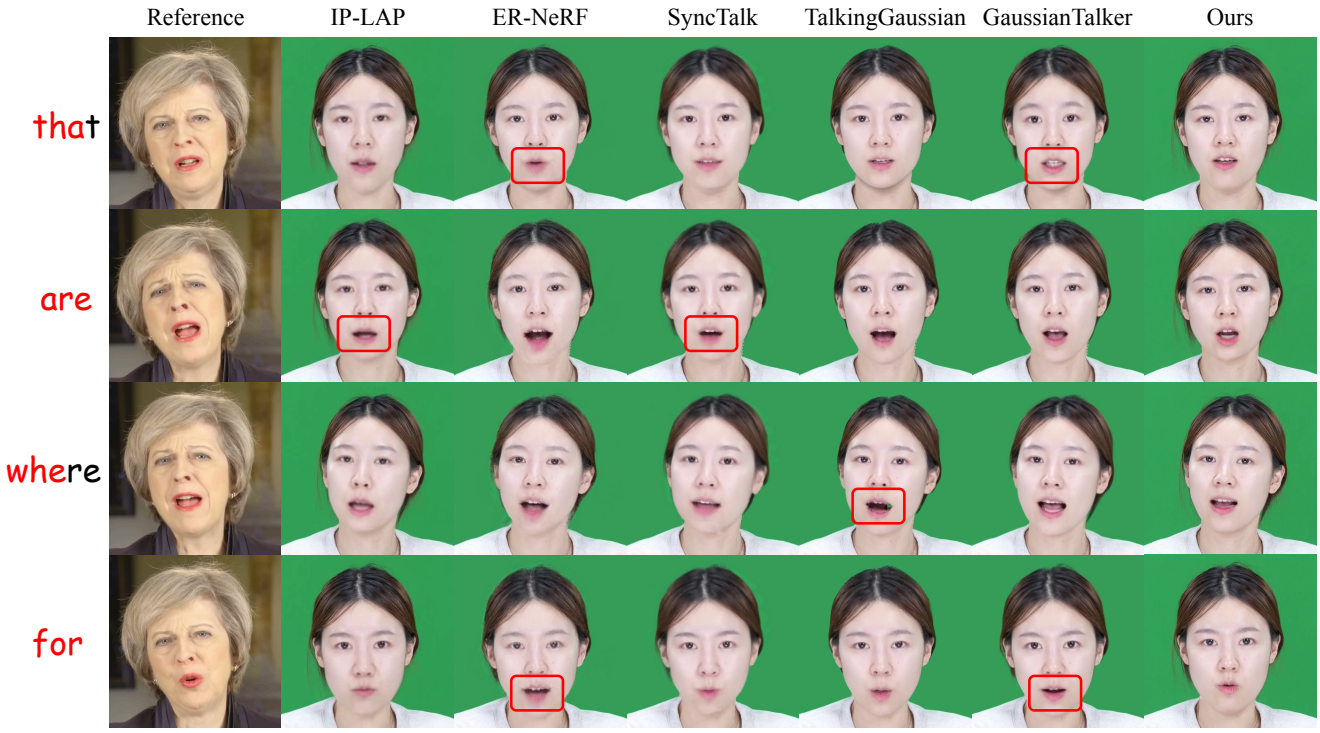


Figure 1. **Qualitative comparison of cross-language reenactment with previous methods.** The training data for this customized character is in Chinese, but we drive her with English during rendering. Our method achieves more precise audio-lip synchronization.

Metrics	Wav2Lip [13]	IP-LAP [23]	ER-NeRF [9]	GeneFace [19]	SyncTalk [12]	TalkingGaussian [10]	GaussianTalker [3]	Ours
Identity Accuracy	2.867	3.041	3.124	2.712	2.421	3.413	<u>3.680</u>	**4.205**
Lip-sync Accuracy	<u>3.796</u>	2.902	2.358	3.018	3.523	3.234	3.354	**4.033**
Head-pose Smoothness	3.487	<u>3.530</u>	2.871	2.955	3.214	2.837	3.249	**3.964**
Image Quality	2.439	3.640	3.196	3.277	3.203	3.651	<u>3.730</u>	**4.026**
Video Realness	3.361	<u>3.624</u>	2.718	2.939	3.094	3.415	3.389	**4.132**

Table 2. **User Study.** Rating is on a scale of 1-5; the higher the better. We highlight the **best** and <u>second-best</u> results. We achieve state-of-the-art performance on all metrics.

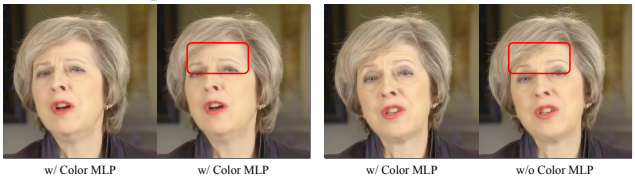


Figure 2. **Qualitative Ablations of color MLP.** Removing the color MLP would result in smoother textures, such as wrinkles being erased.

periments, including 2D-based methods: Wav2Lip [13], DINet [22], IP-LAP [23]; NeRF-based methods: AD-NeRF [5], RAD-NeRF [15], ER-NeRF [9], GeneFace [19], GeneFace++ [18] and SyncTalk [12]; and 3DGS-based methods: TalkingGaussian [10] and GaussianTalker [3].

In the self-reconstruction experiment, we follow the division from previous works [5, 9], splitting each video into training and test sets at a 10:1 ratio. For 2D-based methods, we use the officially released pre-trained models, rendering with silent test video frames and corresponding audio. For 3D-based methods, we train a dedicated model for each identity using the official code. During evaluation, test-set audio is used for rendering.

In the cross-identity experiment, we choose three English-speaking videos — two female and one male speaker. We drive the models with audio from speakers of different genders to ensure the driving audio has a different timbre from the training set. In the cross-language experiment, we render English speakers with French audio and Chinese speakers with English audio.

Additionally, two impressive works are not yet open-sourced [1, 20], preventing detailed comparisons. However, we discuss our principled advantages over them. GaussianTalker[1] [20] lacks the Expression-Visual Prior stage in

[1]There are two concurrent works both named GaussianTalker, and we've already provided a detailed comparison with the other one in the main text and supplementary video.

our method, so we expect it would struggle with large facial motions (as shown in our ablation study when removing Expression-Visual Prior). GaussianSpeech [1] uses a different setup, capturing expensive multi-view synchronized videos for each identity to achieve consistent 3D rendering from different angles. In contrast, we show that learning with the Expression-Visual Prior allows our method to achieve similar 3D consistency with just a few minutes of monocular video of the target person. Furthermore, GaussianSpeech doesn't include Audio-Expression Prior learning, which we believe limits its generalization to OOD audio. For instance, in its official demo, the female speaker on the left exhibits imperfect lip-sync in OOD settings. By comparison, our method requires fewer recording and computational resources for specific identity while delivering superior results.

3. Preliminaries of FLAME

To quantify expressions and poses of driving representation, we use the widely recognized FLAME 3D morphable model (3DMM) [11]. Compared to previous head and facial models, FLAME is more accurate and expressive, while still being compatible with standard graphics software. Additionally, FLAME explicitly models head pose and eye rotation. Specifically, the FLAME model represents the head avatar as follows:

$$TP(\vec{\beta}, \vec{\theta}, \vec{\psi}) = \bar{T} + BS(\vec{\beta}; S) + BP(\vec{\theta}; P) + BE(\vec{\psi}; E) \tag{1}$$

where $\bar{T}$ is the FLAME template mesh. $BS(\vec{\beta}; S)$ is the shape blend-shape function to calculate the facial shape, which is the same for every identity. $BP(\vec{\theta}; P)$ is the jaw and neck blend-shape function to represent the motion at these two hinges. While the expression blend-shape $BE(\vec{\psi}; E)$ is able to portray facial expressions that are either obvious or subtle, such as frowns and grimaces.

4. User Study.

To conduct a more comprehensive subjective analysis of GGTalker compared to previous methods, we design a detailed user survey. We select 24 video clips, each lasting 8-15 seconds. Each method is represented by three clips. The survey was designed using the Mean Opinion Score (MOS) evaluation scheme, where participants are asked to rate the generated videos on five perspectives: (1) Identity Accuracy, (2) Lip-sync Accuracy, (3) Head-pose Smoothness, (4) Image Quality, and (5) Video Realness. We invite 37 volunteers aged between 20 and 60 to provide their ratings, as shown in Tab 2. The user study indicate that our method can generate visually excellent quality as perceived by humans, achieving high realism.

5. Ethical Considerations

Our proposed GGTalker can create realistic talking head from a reference video, with potential applications in digital humans, AR/VR, video conferencing, and film production. However, this technology could also be misused, so we propose several measures to mitigate abuse:

Watermarking We'll implement both visible and invisible watermarks in our code. Visible watermarks will identify synthetic content, while invisible ones will track its origin and modification history, making it harder to misuse.

Advancing Detection Algorithms High-quality avatars generated with modern techniques (e.g., 3DGS, diffusion models) are difficult to detect. We'll collaborate with researchers to enhance deepfake detection methods.

Regulation and Awareness It's essential to raise public awareness about deepfake risks and advocate for legal frameworks to regulate their usage.

By balancing innovation with security, we aim to promote responsible development of deepfake technology.

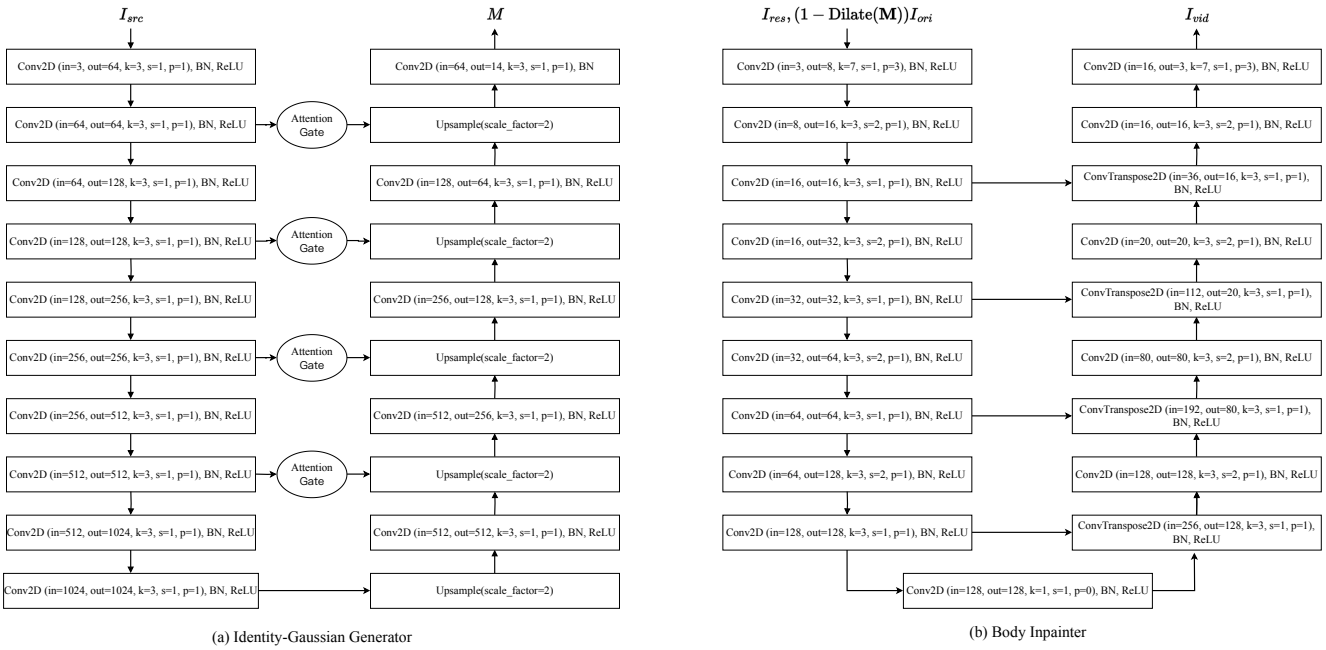
6. Limitations and Future Work

GGTalker has made significant progress in generalizable talking head synthesis, but it still has limitations. First, it focuses solely on synthesizing the head region and cannot generate the torso or background. We hope to find a way to represent the torso and background with a small number of Gaussians, similar to some NeRF-based methods [9, 15]. Second, GGTalker cannot extrapolate emotions. For instance, if the training video captures an angry expression, we can reenact anger. However, if the training video shows a neutral expression, we cannot synthesize anger. In fact, we tried replacing our Audio-Expression module with emotion-controllable expression synthesis methods like DeepTalk [7] and ProbTalk3D [16], but these attempts failed — the synthesized characters showed no obvious emotional changes. We believe that emotional expressiveness relies not only on facial expressions but also on texture changes. For example, anger may cause furrowed brows, while happiness smooths out wrinkles. However, due to the lack of large-scale emotional talking head datasets with diverse identities, predicting facial textures under different emotions remains challenging. We leave this potential contribution to future researchers.

References

[1] Shivangi Aneja, Artem Sevastopolsky, Tobias Kirschstein, Justus Thies, Angela Dai, and Matthias Nießner. Gaussianspeech: Audio-driven gaussian avatars. *arXiv preprint arXiv:2411.18675*, 2024. 2, 3

[2] Chen Chen, Dong Wang, and Thomas Fang Zheng. Cn-cvs: A mandarin audio-visual dataset for large vocabulary continuous visual to speech synthesis. In *ICASSP 2023-2023 IEEE*



(a) Identity-Gaussian Generator

(b) Body Inpainter

Figure 3. **Network architectures** of our Identity-Gaussian Generator and Body Inpainter. In each convolutional layer, "k", "s", and "p" refer to kernel size, stride, and padding, respectively. BN stands for batch normalization.

International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE, 2023. 1

[3] 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 *ACM Multimedia 2024*, 2024. 2

[4] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022. 1

[5] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021. 2

[6] 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. 1

[7] Jisoo Kim, Jungbin Cho, Joonho Park, Soonmin Hwang, Da Eun Kim, Geon Kim, and Youngjae Yu. Deeptalk: Dynamic emotion embedding for probabilistic speech-driven 3d face animation. *arXiv preprint arXiv:2408.06010*, 2024. 3

[8] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 42(4), 2023. 1

[9] 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*, pages 7568–7578, 2023. 2, 3

[10] 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. *arXiv preprint arXiv:2404.15264*, 2024. 2

[11] 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. 1, 3

[12] Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Jun He, Hongyan Liu, and Zhaoxin Fan. Synctalk: The devil is in the synchronization for talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 666–676, 2024. 2

[13] 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 *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492, 2020. 2

[14] Shenhan Qian. Vhap: Versatile head alignment with adaptive appearance priors, 2024. 1

[15] Jiaxiang Tang, Kaisiyuan Wang, Hang Zhou, Xiaokang Chen, Dongliang He, Tianshu Hu, Jingtuo Liu, Gang Zeng, and Jingdong Wang. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. *arXiv preprint arXiv:2211.12368*, 2022. 2, 3

[16] Sichun Wu, Kazi Injamamul Haque, and Zerrin Yumak. Probtalk3d: Non-deterministic emotion controllable speech-driven 3d facial animation synthesis using vq-vae. In *Proceedings of the 17th ACM SIGGRAPH Conference on Motion, Interaction, and Games*, pages 1–12, 2024. 3

[17] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *The IEEE Con-*

ference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022. 1

[18] Zhenhui Ye, Jinzheng He, Ziyue Jiang, Rongjie Huang, Jiawei Huang, Jinglin Liu, Yi Ren, Xiang Yin, Zejun Ma, and Zhou Zhao. Geneface++: Generalized and stable real-time audio-driven 3d talking face generation. *arXiv preprint arXiv:2305.00787*, 2023. 2

[19] 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. 2

[20] 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

[21] 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. 1

[22] Zhimeng Zhang, Zhipeng Hu, Wenjin Deng, Changjie Fan, Tangjie Lv, and Yu Ding. Dinet: Deformation inpainting network for realistic face visually dubbing on high resolution video. *arXiv preprint arXiv:2303.03988*, 2023. 2

[23] 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. 2