HifiDiff: High-Fidelity Talking Face Video Synthesis via Conditional Diffusion

Van-Thien Phan[1], Hyung-Jeong Yang[1*], Seungwon Kim[1], Ji-Eun Shin[2], Soo-Hyung Kim[1]
[1]Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, Korea
[2]Department of Psychology, Chonnam National University, Gwangju, Korea
*Corresponding Author
phanthienbka@gmail.com, {hjyang, seungwon.kim, jieunshin, shkim}@jnu.ac.kr

Abstract

Rapid advances in AI have enabled realistic virtual characters with applications in gaming, education, and entertainment. Audio-driven talking face generation remains challenging due to issues in image fidelity and natural blinking. We propose HifiDiff, a conditional diffusion framework that takes reference frames, masked faces, audio, and blinking cues to produce high-quality avatars. Our method achieves enhanced visual realism, accurate lip-sync, and natural eye blinking, outperforming prior approaches. Extensive experiments on CREMA-D demonstrate its effectiveness, highlighting its potential to advance expressive and controllable talking face generation.

1. Introduction

Talking face generation seeks to synthesize realistic facial movements from speech, with emphasis on lip synchronization, facial coherence, and visual quality. Despite recent progress, producing natural and expressive results remains challenging due to the complexity of human expressions and speech dynamics.

Early works [1, 2] mainly employed GANs, but adversarial training often causes mode collapse, temporal artifacts, and unstable performance, especially at higher resolutions. Recent diffusion-based methods [3, 4, 5] improve fidelity but rely on motion frames, leading to quality degradation in long sequences. In contrast, 3D-based approaches [6] better preserve geometry and coherence, yet typically require identity-specific training, limiting scalability.

We propose a pixel-space conditional diffusion model for generalized talking face synthesis. Unlike landmark-based diffusion methods that often yield misaligned expressions, our framework employs a lip-sync discriminator to learn implicit audio- lip correspondence, while explicitly controlling pose and eye blinking through structured signals. A landmark-based masking strategy further improves fidelity by focusing generation on the facial region, and frame interpolation enhances temporal coherence for smoother videos.

Our main contributions are as follows: (1) A novel pixel-space diffusion framework for high-quality talking face generation. (2) A landmark-constrained masking mechanism enabling accurate lip sync and natural blinking with identity preservation. (3) Long-duration synthesis with consistent control over lip, eye, pose, and facial dynamics.

2. Related Work

Talking head generation seeks not only accurate lip- speech alignment but also natural eye blinking and head motion. 3D-based methods [7, 8] using 3D Morphable Models or Neural Radiance Fields [9, 10] improve realism but often require identity-specific training, limiting generalization.

2D GAN-based approaches [11] have advanced visual quality, yet suffer from instability and mode collapse [9]. Landmark-driven methods [14] add structure but still fail in achieving precise lip synchronization.

Recently, diffusion models have emerged as more stable and expressive. DiffTalk [12] enhances fidelity but struggles with cross-identity lip-sync; Diff2Lip [13] achieves accurate lips but neglects blinking and pose; Diffused Heads [30, 5] mitigate motion degradation but remain limited in long sequences.

Our framework addresses these issues by unifying lip-sync, head pose, and blinking within a diffusion-based model for high-fidelity, temporally coherent talking head generation.

3. Method

In this section, we provide a detailed description of the proposed method and its components.
A framework for inference in Fig 1., diffusion model conditioned on masked frames, source identity, audio features, and blinking signals.

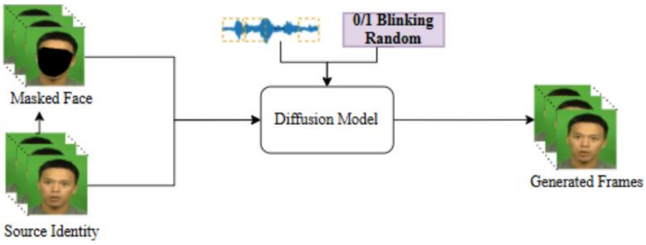

Figure 1: Overview of the proposed talking head synthesis framework using a Diffusion Model for inference.

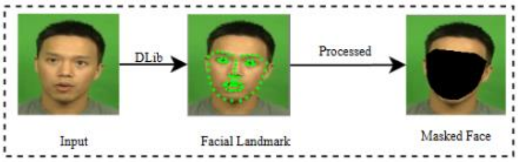

Figure 2: Facial landmark detection and masking. DLib detects 68 landmarks to create a facial region mask, producing a masked image for further processing.

Masked Face. Fig. 2 illustrates masked face generation. We employ DLib [15] for robust real-time landmark detection, which handles facial rotation effectively and enables reliable preprocessing.

Eye Blinking. Blink signals are detected using the Eye Aspect Ratio (EAR) for both eyes, following Zhang et al. [16], to quantify eye openness.

Objective Function. For training, our model uses the loss framework to ensure realism and lip-sync, with each loss term optimizing a specific aspect of talking face generation. In the denoising function $\varepsilon_\theta(x_t, t)$ [17] estimates the noise component present in the corrupted sample x_t. The model is optimized using a simple mean squared error (MSE) loss, defined as:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,\mathbf{x}_0,\varepsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\varepsilon - \varepsilon_\theta(\mathbf{x}_t, t)\|_2^2\right] \quad (1)$$

Following [15], we compute $x_\theta(x_t, t)$ as an estimate of the clean frame x_0. An L_2 reconstruction loss ensures matches x0 at each timestep, improving denoising accuracy:

$$\mathcal{L}_{L2} = \mathbb{E}_{x_{0,s},t,\varepsilon}\left[\left\|x_{0,s}^\theta - x_{0,s}\right\|_2^2\right] \quad (2)$$

The model aligns lip movements with audio by minimizing the cosine similarity between video embeddings x and audio embeddings a using a pre-trained SyncNet [18]:

$$\mathcal{L}_{\text{sync}} = \mathbb{E}_{x_{0,s},t,\varepsilon}\left[\text{SyncNet}\left(x_{0\,s:s+5}^\theta, a_{s:s+5}\right)\right] \quad (3)$$

To preserve visual details, a perceptual loss minimizes differences between generated and reference frame features using VGG-19 [19]:

$$\mathcal{L}_{\text{Lpips}} = \mathbb{E}_{x_{0,s},t,\varepsilon}\mathbb{E}_l\left[\left\|\phi_l(x_{0\,s}^\theta) - \phi_l(x_{0,s})\right\|_2^2\right] \quad (4)$$

The total loss is a weighted sum:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{simple}} + \lambda_2\mathcal{L}_{\text{L2}} + \lambda_3\mathcal{L}_{\text{sync}} + \lambda_4\mathcal{L}_{\text{Lpips}} \quad (5)$$

with $\lambda 1=1$, $\lambda 2=1$, $\lambda 3=0.15$, $\lambda 4=0.25$.

4. Experiment

4.1 Setup

Dataset. We evaluate our method on the CREMA-D dataset [26], containing 7,442 clips from 91 speakers, each 1-5 seconds long, capturing diverse emotions. Following Stypulkowski et al. [4], we split the dataset into 90% for training and 10% for testing. Videos are resized to 128×128 pixels at 25fps, and audio is resampled to 16kHz and converted into spectrograms (16×80) using STFT (window=800, hop=200).

Implementation. Our model uses a UNet-based architecture [4] with an initial channel size of 128 and channel multipliers (1, 2, 2, 4). An audio encoder extracts high-level features, and the diffusion process employs T=1000 timesteps for training, reduced to T=50 at inference. Frame interpolation [17] mitigates temporal jitter.

Evaluation. We assess visual quality using FID [20] and FVD [21], lip-sync accuracy via LSEC [22], and eye-blinking realism through blink frequency and duration [4].

4.2 Comparision with SOTA methods

We compare our method with several state-of-the-art approaches: SDA [25], MakeItTalk [14], Wav2Lip [24], PC-AVS [23], EAMM [9], and Diffused Heads [4]. Fig. 3 shows visual comparisons on CREMA-D. Our model consistently produces high-fidelity results with natural pose, accurate lip-sync, and realistic eye blinking (Table 1).

GAN-based methods like SDA often generate artifacts and reduce realism. MakeItTalk, relying on facial landmarks, shows poor audio- face synchronization, while Wav2Lip achieves good lip-sync but unnatural pose and blinking. PC-AVS, using modulated convolutions, introduces image artifacts, and EAMM, despite modeling audio-to-facial dynamics, produces blur and imperfect lip alignment. Diffused Heads improves blinking, head motion, and image quality but struggles with long-sequence generation, where outputs can collapse over time.

Method	FVD ↓	FID ↓	Blinks/s	Blink Duration	LSE_C ↑
GT	-	-	0.24	0.4	5.88
SDA	376.48	79.82	**0.25**	0.26	5.10
MakeItTalk	256.88	17.26	0.02	0.80	3.71
Wav2Lip	193.32	12.57	-	-	**6.08**
PC-AVS	333.94	22.53	0.02	0.20	5.67
EAMM	196.82	19.40	-	-	4.22
Diffused Heads	88.614	12.45	0.28	**0.36**	4.56
Our	**74.68**	**11.51**	0.30	0.27	5.73

Table 1: Results evaluating the impact of different methods. ↓: lower is better, ↑: higher is better.

4.3 Ablation Studies

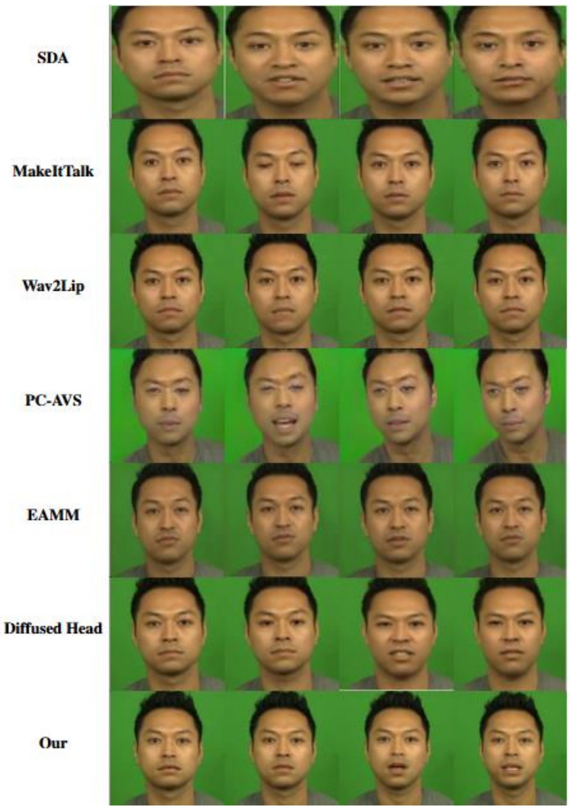

Figure 3: Qualitative comparison between existing methods and our proposed approach. Each row corresponds to a different method.

We conduct ablation experiments to assess the contribution of four input components: masked face, eye-blinking signal, audio condition, and reference frames. Three settings are compared: (1) all inputs except eye-blinking, (2) full model, and (3) all inputs except masked face. Each variant is evaluated on visual quality, lip-sync accuracy, blink rate, and median blink duration. Results (Table 5) show that the full model achieves the best balance, closely matching ground truth and demonstrating the importance of each input modality.

Method	FVD ↓	FID ↓	Blinks/s	Blink Duration	LSE_C ↑
GT	-	-	0.24	0.40	5.88
Set 1 (w/o blinking)	**72.29**	**11.39**	-	-	**5.82**
Set 2 (full input)	74.68	11.51	**0.30**	**0.27**	5.73
Set 3 (w/o masked face)	487.93	18.06	1.00	0.12	4.58

Table 2: Ablation study of different input configurations.

5. Conclusion

We present a diffusion-based framework for talking face generation that integrates masked face guidance, audio conditions, eye-blinking signals, and reference frames. Our approach achieves accurate lip synchronization, natural pose and blinking, and high visual fidelity while preserving speaker identity. Quantitative and qualitative results on CREMA-D show that our method outperforms prior state-of-the-art approaches, enabling more expressive and controllable talking face generation.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00219107, %). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629, %) grant funded by the Korea government (MSIT). This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government (MSIT)(IITP-2025-RS-2022-00156287, %).

Reference

[1] Jian Guo, Dongdong Zhang, Xiaoyang Liu, Zhongang Zhang, Yifan Zhang, Ping Luo, and Deli Zhao. Liveportrait: Efficient portrait animation with stitching and retargeting control. arXiv preprint arXiv:2407.03168, 2024.

[2] Fang-Ting Hong, Liang Zhang, Lin Shen, and Dong Xu. Depth-aware generative adversarial network for talking head video generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3397–3406, 2022.

[3] Dragos Bigioi, Subhrajyoti Basak, Heather Jordan, Rachel McDonnell, and Peter Corcoran. Speech driven video editing via an audio-conditioned diffusion model. arXiv preprint arXiv:2301.04474, 2023.

[4] Michał Stypułkowski, Konstantinos Vougioukas, Sen He, Maciej Zi˛eba, Stavros Petridis, and Maja Pantic. Diffused heads: Diffusion models beat gans on talking face generation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 5091–5100, 2024.

[5] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. arXiv preprint arXiv:2402.17485, 2024.

[6] Shunyu Yao, RuiZhe Zhong, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. Dfanerf: Personalized talking head generation via disentangled face attributes neural rendering. arXiv preprint arXiv:2211.14108, 2022.

[7] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[8] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In Proceedings of the European Conference on Computer Vision (ECCV), 2020.

[9] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-aware implicit neural audio-driven video portrait generation. arXiv preprint arXiv:2204.11888, 2022.

[10] Shuai Shen, Wanhua Li, Zheng Zhu, Yueqi Duan, Jie Zhou, and Jiwen Lu. Learning dynamic facial radiance fields for few-shot talking head synthesis. In Proceedings of the European Conference on Computer Vision (ECCV), 2022.

[11] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. International Journal of Computer Vision, 128:1398–1413, 2020.

[12] Sourya Mukhopadhyay, Srijan Suri, Ravi Teja Gadde, and Abhinav Shrivastava. Diff2lip: Audio conditioned diffusion models for lip-synchronization. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 5280–5290, 2024. doi: 10.1109/WACV57701.2024.00521.

[13] Shuhang Shen et al. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1982–1991, 2023. doi: 10.1109/CVPR52729.2023.00197.

[14] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makeittalk: Speaker-aware talking-head animation. ACM Transactions on Graphics (TOG), 39(6):1–15, 2020.

[15] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1867–1874, 2014. doi: 10.1109/CVPR.2014.241.

[16] 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/CV Conference on Computer Vision and Pattern Recognition (CVPR), pages 8652–8661, 2023.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems (NeurIPS), volume 33, pages 6840–6851, 2020

[18] KR 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 ACM International Conference on Multimedia (ACMMM), 2020.

[19] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 586–595, 2018.

[20] 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. In Advances in Neural Information Processing Systems, volume 30, 2017.

[21] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. arXiv preprint arXiv:1812.01717, 2019.

[22] Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In Computer Vision – ACCV 2016 Workshops, pages 251–263. Springer, 2017.

[23] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and ZiweiLiu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 4176–4186, 2021.

[24] Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and C.V. Jawahar. Towards automatic face-to-face translation. In Proceedings of the ACM International Conference on Multimedia (ACMMM), 2019.

[25] Kuangxiao Gu, Yuqian Zhou, and Thomas Huang. Flnet: Landmark driven fetching and learning network for faithful talking facial animation synthesis. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2020.

[26] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing, 5(4):377–390, 2014.