This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DiffTalk: Crafting Diffusion Models for Generalized Audio-Driven Portraits Animation

Shuai Shen[1] Wenliang Zhao[1] Zibin Meng[1] Wanhua Li[1] Zheng Zhu[2] Jie Zhou[1] Jiwen Lu[1,*]
[1]Tsinghua University [2]PhiGent Robotics

Figure 1. We present a crafted conditional **Diff**usion model for generalized **Talk**ing head synthesis (DiffTalk). Given a driven audio, the DiffTalk is capable of synthesizing high-fidelity and synchronized talking videos for multiple identities without further fine-tuning.

Abstract

Talking head synthesis is a promising approach for the video production industry. Recently, a lot of effort has been devoted in this research area to improve the ***generation quality*** *or enhance the* ***model generalization****. However, there are few works able to address both issues simultaneously, which is essential for practical applications. To this end, in this paper, we turn attention to the emerging powerful Latent* ***Diff****usion Models, and model the* ***Talk****ing head generation as an audio-driven temporally coherent denoising process (DiffTalk). More specifically, instead of employing audio signals as the single driving factor, we investigate the control mechanism of the talking face, and incorporate reference face images and landmarks as conditions for personality-aware generalized synthesis. In this way, the proposed DiffTalk is capable of producing high-quality talking head videos in synchronization with the source audio, and more importantly, it can be naturally generalized across different identities without further fine-tuning. Additionally, our DiffTalk can be gracefully tailored for higher-resolution synthesis with negligible extra computational cost. Extensive experiments show that the proposed DiffTalk efficiently synthesizes high-fidelity audio-driven talking head videos for generalized novel identities. For more video results, please refer to* https://sstzal.github.io/DiffTalk/.

*Corresponding author

1. Introduction

Talking head synthesis is a challenging and promising research topic, which aims to generate video portraits with given audio. This technique is widely applied in various practical scenarios including animation, virtual avatars, online education, and video conferencing [4, 45, 48, 51, 54].

Recently a lot of effort has been devoted to this research area to improve the **generation quality** or enhance the **model generalization**. Among these existing main-

stream talking head generation approaches, the 2D-based methods usually depend on generative adversarial networks (GANs) [6, 10, 16, 23, 29] for audio-to-lip mapping, and most of them perform competently on model generalization. However, since GANs need to simultaneously optimize a generator and a discriminator, the training process lacks stability and is prone to mode collapse [11]. Due to this restriction, the generated talking videos are of limited image quality, and difficult to scale to higher resolutions. By contrast, 3D-based methods [2, 17, 43, 47, 55] perform better in synthesizing higher-quality talking videos. Whereas, they highly rely on identity-specific training, and thus cannot generalize across different persons. Such identity-specific training also brings heavy resource consumption and is not friendly to practical applications. Most recently, there are some 3D-based works [37] that take a step towards improving the generalization of the model. However, further fine-tuning on specific identities is still inevitable.

Generation quality and model generalization are two essential factors for better deployment of the talking head synthesis technique to real-world applications. However, few existing works are able to address both issues well. In this paper, we propose a crafted conditional **Diff**usion model for generalized **Talk**ing head synthesis (DiffTalk), that aims to tackle these two challenges simultaneously. Specifically, to avoid the unstable training of GANs, we turn attention to the recently developed generative technology Latent Diffusion Models [31], and model the talking head synthesis as an audio-driven temporally coherent denoising process. On this basis, instead of utilizing audio signals as the single driving factor to learn the audio-to-lip translation, we further incorporate reference face images and landmarks as supplementary conditions to guide the face identity and head pose for personality-aware video synthesis. Under these designs, the talking head generation process is more controllable, which enables the learned model to naturally generalize across different identities without further fine-tuning. As shown in Figure 1, with a sequence of driven audio, our DiffTalk is capable of producing natural talking videos of different identities based on the corresponding reference videos. Moreover, benefiting from the latent space learning mode, our DiffTalk can be gracefully tailored for higher-resolution synthesis with negligible extra computational cost, which is meaningful for improving the generation quality.

Extensive experiments show that our DiffTalk can synthesize high-fidelity talking videos for novel identities without any further fine-tuning. Figure 1 shows the generated talking sequences with one driven audio across three different identities. Comprehensive method comparisons show the superiority of the proposed DiffTalk, which provides a strong baseline for the high-performance talking head synthesis. To summarize, we make the following contributions:

- We propose a crafted conditional diffusion model for high-quality and generalized talking head synthesis. By introducing smooth audio signals as a condition, we model the generation as an audio-driven temporally coherent denoising process.
- For personality-aware generalized synthesis, we further incorporate dual reference images as conditions. In this way, the trained model can be generalized across different identities without further fine-tuning.
- The proposed DiffTalk can generate high-fidelity and vivid talking videos for generalized identities. In experiment, our DiffTalk significantly outperforms 2D-based methods in the generated image quality, while surpassing 3D-based works in the model generalization ability.

2. Related Work

Audio-driven Talking Head Synthesis. The talking head synthesis aims to generate talking videos with lip movements synchronized with the driving audio [14, 41, 53]. In terms of the modeling approach, we roughly divide the existing methods into 2D-based and 3D-based ones. In the 2D-based methods, GANs [6, 10, 16, 29] are usually employed as the core technologies for learning the audio-to-lip translation. Zhou *et al.* [54] introduce a speaker-aware audio encoder for personalized head motion modeling. Prajwal *et al.* [29] boost the lip-visual synchronization with a well-trained Lip-Sync expert [8]. However, since the training process of GANs lacks stability and is prone to mode collapse [11], the generated talking videos are always of limited image quality, and difficult to scale to higher resolutions. Recently a series of 3D-based methods [4, 21, 40–42] have been developed. [40–42] utilize 3D Morphable Models [2] for parametric control of the talking face. More recently, the emerging Neural radiance fields [27] provide a new solution for 3D-aware talking head synthesis [3, 17, 25, 37]. However, most of these 3D-based works highly rely on identity-specific training, and thus cannot generalize across different identities. Shen *et al.* [37] have tried to improve the generalization of the model, however, further fine-tuning on specific identities is still inevitable. In this work, we propose a brand-new diffusion model-based framework for high-fidelity and generalized talking head synthesis.

Latent Diffusion Models. Diffusion Probabilistic Models (DM) [38] have shown strong ability in various image generation tasks [11, 19, 30]. However, due to pixel space-based training [31, 33], very high computational costs are inevitable. More recently, Rombach *et al.* [31] propose the Latent Diffusion Models (LDMs), and transfer the training and inference processes of DM to a compressed lower-dimension latent space for more efficient computing [13, 50]. With the democratizing of this technology, it has been successfully employed in a series of works, in-

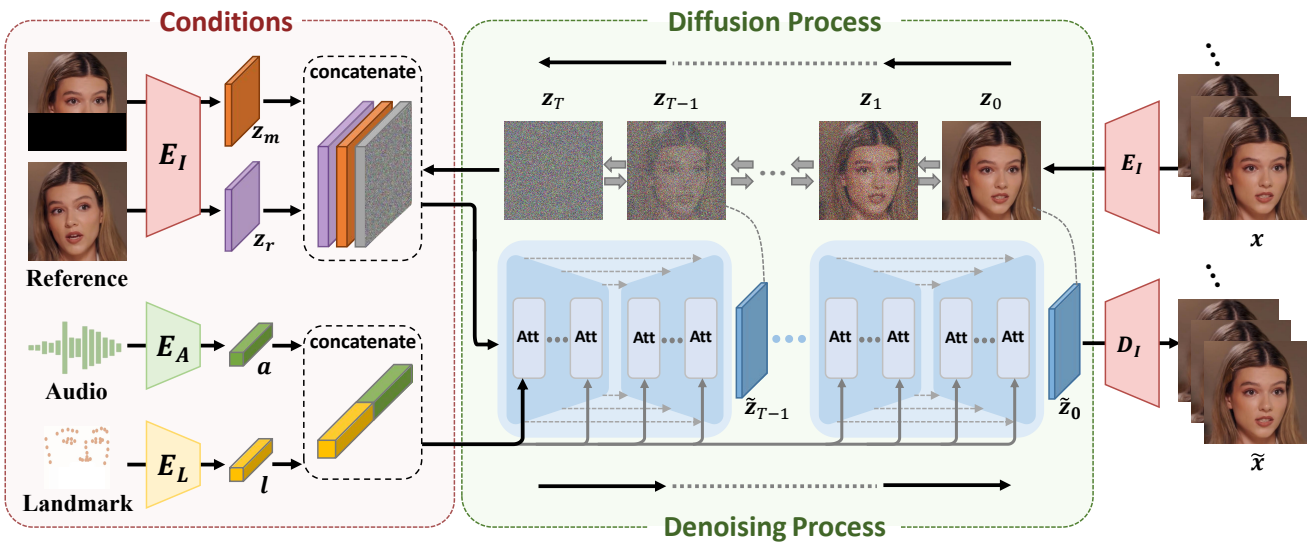

Figure 2. Overview of the proposed DiffTalk for generalized talking head synthesis. Apart from the audio condition to drive the lip motions, we further incorporate reference images and facial landmarks as extra driving factors for personalized facial modeling. In this way, the generation process is more controllable, which enables the learned model to generalize across different identities without further fine-tuning. Furthermore, we can gracefully improve our DiffTalk for higher-resolution synthesis with slight extra computational cost.

cluding text-to-image translation [22, 32, 34], super resolution [7, 12, 28], image inpainting [24, 26], motion generation [36, 49], 3D-aware prediction [1, 35, 44]. In this work, drawing on these successful practices, we model the talking head synthesis as an audio-driven temporally coherent denoising process and achieve superior generation results.

3. Methodology

3.1. Overview

To tackle the challenges of generation quality and model generalization, we model the talking head synthesis as an audio-driven temporally coherent denoising process, and term the proposed method as DiffTalk. An overview of the DiffTalk is shown in Figure 2. By introducing smooth audio features as a condition, we improve the diffusion model for temporally coherent facial motion modeling. For further personalized facial modeling, we incorporate reference face images and facial landmarks as extra driving factors. In this way, the talking head generation process is more controllable, which enables the learned model to generalize across different identities without any further fine-tuning. Moreover, benefiting from the latent space learning mode, we can graceful improve our DiffTalk for higher-resolution synthesis with negligible extra computational cost, which contributes to improving the generation quality.

3.2. Conditional Diffusion Model for Talking Head

The emergence of Latent Diffusion Models (LDMs) [19, 31] provides a straightforward and effective way for high-fidelity image synthesis. To inherit its excellent properties, we adopt this advanced technology as the foundation of our method and explore its potential in modeling the dynamic talking head. With a pair of well-trained image encoder E_I and decoder D_I which are frozen in training [13], the input face image $x \in \mathbb{R}^{H\times W\times 3}$ can be encoded into a latent space $z_0 = E_I(x) \in \mathbb{R}^{h\times w\times 3}$, where $H/h = W/w = f$, H, W are the height and width of the original image and f is the downsampling factor. In this way, the learning is transferred to a lower-dimensional latent space, which is more efficient with fewer train resources. On this basis, the standard LDMs are modeled as a time-conditional UNet-based [33] denoising network $\mathcal{M}$, which learns the reverse process of a Markov Chain [15] of length T. The corresponding objective can be formulated as:

$$L_{LDM} := \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \mathcal{M}\left(z_t, t\right)\|_2^2\right], \qquad (1)$$

where $t \in [1, \cdots, T]$ and z_t is obtained through the forward diffusion process from z_0. $\tilde{z}_{t-1} = z_t - \mathcal{M}(z_t, t)$ is the denoising result of z_t at time step t. The final denoised result $\tilde{z}_0$ is then upsampled to the pixel space with the pre-trained image decoder $\tilde{x} = D_I(\tilde{z}_0)$, where $\tilde{x} \in \mathbb{R}^{H\times W\times 3}$ is the reconstructed face image.

Given a source identity and driven audio, our goal is to train a model for generating a natural target talking video in synchronization with the audio condition while maintaining the original identity information. Furthermore, the trained model also needs to work for novel identities during inference. To this end, the audio signal is introduced as a basic condition to guide the direction of the denoising process for

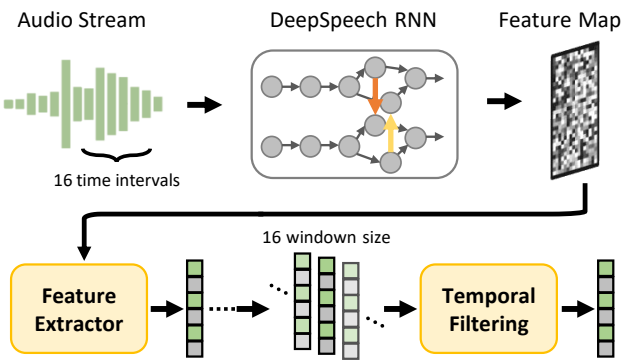

Figure 3. Visualization of the smooth audio feature extractor.

modeling the audio-to-lip translation.

Smooth Audio Feature Extraction. To better incorporate temporal information, we involve two-stage smoothing operations in the audio encoder E_A, as shown in Figure 3. Firstly, following the practice in VOCA [9], we reorganize the raw audio signal into overlapped windows of size 16 time intervals (corresponding to audio clips of 20ms), where each window is centered on the corresponding video frame. A pre-trained RNN-based DeepSpeech [18] module is then leveraged to extract the per-frame audio feature map F. For better inter-frame consistency, we further introduce a learnable temporal filtering [42]. It receives a sequence of adjacent audio features $[F_{i-w}, \ldots, F_i, \ldots, F_{i+w}]$ with $w = 8$ as input, and computes the final smoothed audio feature for the i-th frame as $a \in \mathbb{R}^{D_A}$ in a self-attention-based learning manner, where D_A denotes the audio feature dimension. By encoding the audio information, we bridge the modality gap between the audio signals and the visual information. Introducing such smooth audio features as a condition, we extend the diffusion model for temporal coherence-aware modeling of face dynamics when talking. The objective is then formulated as:

$$L_A := \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,1),a,t}\left[\|\epsilon - \mathcal{M}\left(z_t, t, a\right)\|_2^2\right]. \quad (2)$$

Identity-Preserving Model Generalization. In addition to learning the audio-to-lip translation, another essential task is to realize the model generalization while preserving complete identity information in the source image. Generalized identity information includes face appearance, head pose, and image background. To this end, a reference mechanism is designed to empower our model to generalize to new individuals unseen in training, as shown in Figure 2. Specifically, a random face image x_r of the source identity is chosen as a reference condition, which contains appearance and background information. To prevent training shortcuts, we limit the selection of x_r to 60 frames beyond the target image. However, since the ground-truth face image has a completely different pose from x_r, the model is expected to transfer the pose of x_r to the target face without any prior information. This is somehow an ill-posed problem with no unique solution. For this reason, we further incorporate the masked ground-truth image x_m as another reference condition to provide the target head pose guidance. The mouth region of x_m is completely masked to ensure that the ground truth lip movements are not visible to the network. In this way, the reference x_r focuses on affording mouth appearance information, which additionally reduces the training difficulty. Before serving as conditions, x_r and x_m are also encoded into the latent space through the trained image encoder, and we have $z_r = D_I(x_r) \in \mathbb{R}^{h\times w\times 3}$, $z_m = D_I(x_m) \in \mathbb{R}^{h\times w\times 3}$. On this basis, an auxiliary facial landmarks condition is also included for better control of the face outline. Similarly, landmarks in the mouth area are masked to avoid shortcuts. The landmark feature $l \in \mathbb{R}^{D_L}$ is obtained with an MLP-based encoder E_L, where D_L is the landmark feature dimension. In this way, combining these conditions with audio feature a, we realize the precise control over all key elements of a dynamic talking face. With $C = \{a, z_r, z_m, l\}$ denoting the condition set, the talking head synthesis is finally modeled as a conditional denoising process optimized with the following objective:

$$L := \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,1),C,t}\left[\|\epsilon - \mathcal{M}\left(z_t, t, C\right)\|_2^2\right], \quad (3)$$

where the network parameters of $\mathcal{M}$, E_A and E_L are jointly optimized via this equation.

Conditioning Mechanisms. Based on the modeling of the conditional denoising process in Eq. 3, we pass these conditions C to the network in the manner shown in Figure 2. Specifically, following [31], we implement the UNet-based backbone $\mathcal{M}$ with the cross-attention mechanism for better multimodality learning. The spatially aligned references z_r and z_m are concatenated channel-wise with the noisy map z_T to produce a joint visual condition $C_v = [z_T; z_m; z_r] \in \mathbb{R}^{h\times w\times 9}$. C_v is fed to the first layer of the network to directly guide the output face in an image-to-image translation fashion. Additionally, the driven-audio feature a and the landmark representation l are concatenated into a latent condition $C_l = [a; l] \in \mathbb{R}^{D_A+D_L}$, which serves as the *key* and *value* for the intermediate cross-attention layers of $\mathcal{M}$. To this extent, all condition information $C = \{C_v, C_l\}$ are properly integrated into the denoising network $\mathcal{M}$ to guide the talking head generation process.

Higher-Resolution Talking Head Synthesis Our proposed DiffTalk can also be gracefully extended for higher-resolution talking head synthesis with negligible extra computational cost and faithful reconstruction effects. Specifically, considering the trade-off between the perceptual loss and the compression rate, for training images of size $256 \times 256 \times 3$, we set the downsampling factor as $f = 4$

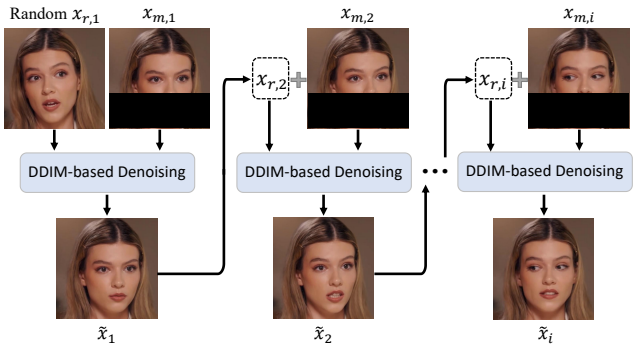

Figure 4. Illustration of the progressive inference strategy.

and obtain the latent space of $64 \times 64 \times 3$. Furthermore, for higher-resolution generation of $512 \times 512 \times 3$, we just need to adjust the paired image encoder E_I and decoder D_I with a bigger downsampling factor $f = 8$. Then the trained encoder is frozen and employed to transfer the training process to a $64 \times 64 \times 3$ latent space as well. This helps to relieve the pressure on insufficient resources, and therefore our model can be gracefully improved for higher-resolution talking head video synthesis.

3.3. Progressive Inference

We perform inference with Denoising Diffusion Implicit Model-based (DDIM) [39] iterative denoising steps to accelerate sampling for more efficient synthesis. To further boost the coherence of the generated talking videos, we develop a progressive reference strategy in the reference process as shown in Figure 4. Specifically, when rendering a talking video sequence with the trained model, for the first frame, $x_{r,1}$ is a random face image from the target identity. Subsequently, the synthetic face image $\tilde{x}_i$ is exploited as the reference $x_{r,i+1}$ for the next frame. In this way, image details between adjacent frames remain consistent, resulting in a smoother transition between frames. It is worth noting that this strategy is not used for training. Since the difference between adjacent frames is small, we need to eliminate such references to avoid learning shortcuts. Following the practice in [31], masked z_T is used during inference, where the mouth area is masked and randomly initialized, allowing the network to focus on the denoising of this region. To further alleviate the video jitter issue, we utilize [20] for frame interpolation to get smoother synthesized talking videos.

4. Experiments

4.1. Experimental Settings

Dataset. To train the audio-driven diffusion model, an audio-visual dataset HDTF [52] is used. It contains 16 hours of talking videos in 720P or 1080P from more than 300 identities. We randomly select 100 videos, and finally form a video gallery with the length of 100 minutes for training, while the remaining data serve as the test set.

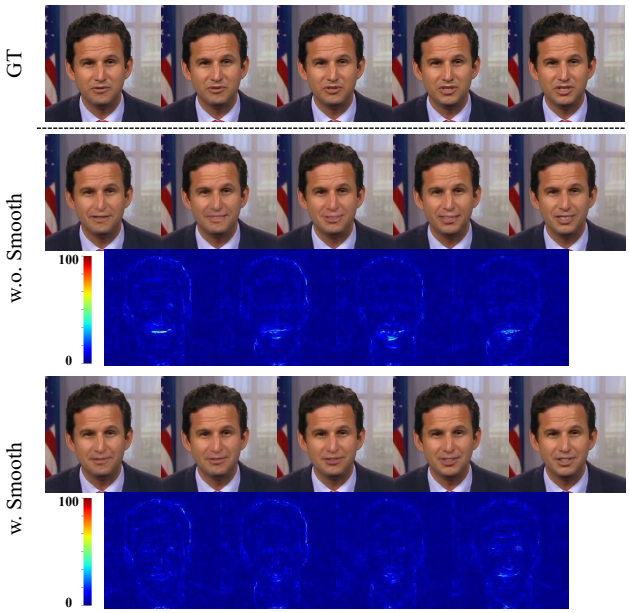

Figure 5. Ablation study on the audio smoothing operation. We show the differences between adjacent frames as heatmaps for better visualization.

Method		PSNR↑	SSIM↑	LPIPS↓	SyncNet↓↑
Test Set A	GT	-	-	-	0/9.610
	w/o	33.67	0.944	**0.024**	1/5.484
	w	**34.17**	**0.946**	**0.024**	**1/6.287**
Test Set B	GT	-	-	-	0/9.553
	w/o	32.70	**0.924**	**0.031**	1/5.197
	w	**32.73**	**0.925**	**0.031**	**1/5.387**

Table 1. Ablation study to investigate the contribution of the audio smoothing operation. 'w' indicates the model is trained with the audio features after temporal filtering and vice versa.

Metric. We evaluate our proposed method through visual results coupled with quantitative indicators. PSNR (↑), SSIM (↑) [46] and LPIPS (↓) [50] are three metrics for assessing image quality. The LPIPS is a learning-based perceptual similarity measure that is more in line with human perception, we therefore recommend this metric as a more objective indicator. The SyncNet score (Offset↓ / Confidence↑) [8] checks the audio-visual synchronization quality, which is important for the audio-driven talking head generation task. ('↓' indicates that the lower the better, while '↑' means that the higher the better.)

Implementation Details. We resize the input image to 256×256 for experiments. The downsampling factor f is set as 4, so the latent space is $64 \times 64 \times 3$. For training the model for higher resolution synthesis, the input is resized to 512×512 with $f = 8$ to keep the same size of latent space.

Figure 6. Ablation study on the design of the conditions. The marks above these images refer to the following meanings, 'A': Audio; 'L': Landmark; 'R': Random reference image; 'M': Masked ground-truth image. We show the generated results under different condition settings on two test sets, and demonstrate the effectiveness of our final design, *i.e.* A+L+M+R.

Method		PSNR↑	SSIM↑	LPIPS↓	SyncNet↓↑
Test Set A	GT	-	-	-	4/7.762
	w/o	**34.17**	**0.946**	0.024	1/6.287
	w	33.95	**0.946**	**0.023**	**-1/6.662**
Test Set B	GT	-	-	-	3/8.947
	w/o	32.73	**0.925**	0.031	1/5.387
	w	**33.02**	**0.925**	**0.030**	**1/5.999**

Table 2. Ablation study on the effect of the progressive inference strategy. 'w/o' indicates that a random reference image is employed as the condition, and 'w' means that the reference is the generated result of the previous frame.

The length of the denoising step T is set as 200 for both the training and inference process. The feature dimensions are $D_A = D_L = 64$. Our model takes about 15 hours to train on 8 NVIDIA 3090 GPUs.

4.2. Ablation Study

Effect of the Smooth Audio. In this subsection, we investigate the effect of the audio smooth operations. Quantitative results in Table 1 show that the model equipped with the audio temporal filtering module outperforms the one without smooth audio, especially in the SyncNet score. We further visualize the differences between adjacent frames as the heatmaps shown in Figure 5. The results without audio filtering present obvious high heat values in the mouth region, which indicates the jitters in this area. By contrast, with smooth audio as the condition, the generated video frames show smoother transitions, which are reflected in the soft differences of adjacent frames.

Design of the Conditions. A major contribution of this work is the ingenious design of the conditions for general and high-fidelity talking head synthesis. In Figure 6, we show the generated results under different condition settings step by step, to demonstrate the superiority of our design. With pure audio as the condition, the model fails to generalize to new identities, and the faces are not aligned with the background in the inpainting-based inference. Adding landmarks as another condition tackles the misalignment problem. A random reference image is further introduced trying to provide the identity information. Whereas, since the ground-truth face image has a different pose from this random reference, the model is expected to transfer the pose of reference to the target face. This greatly increases the difficulty of training, leading to hard network convergence, and the identity information is not well learned. Using the audio and masked ground-truth images as driving factors mitigates the identity inconsistency and misalignment issues, however the appearance of the mouth can not be learned since this information is not visible to the network. For this reason, we employ the random reference face and the masked ground-truth image together for dual driving, where the random reference provides the lip appearance message and the masked ground-truth controls the head pose and identity. Facial landmarks are also incorporated as a condition that helps to model the facial contour better. Results in Figure 6 show the effectiveness of such design in synthesizing realism and controllable face images.

Impact of the Progressive Inference. Temporal correlation inference is developed in this work through the progressive reference strategy. We conduct an ablation study in Table 2 to investigate the impact of this design. 'w/o' indicates that a random reference image x_r is employed, and 'w' means that the generated result of the previous frame is chosen as the reference condition. With such progressive inference, the SyncNet scores are further boosted, since the temporal correlation is better modeled and the talking style

Figure 7. Visual comparison with some representative 2D-based talking head generation methods ATVGnet [5], MakeitTalk [54] and Wav2Lip [29], and with some recent 3D-based ones AD-NeRF [17] and DFRF [37]. The results of DFRF are synthesized with the base model without fine-tuning for fair comparisons. AD-NeRF is trained on these two identities respectively to produce the results.

becomes more coherent. The LPIPS indicator is also enhanced with this improvement. PSNR tends to give higher scores to blurry images [50], so we recommend LPIPS as a more representative metric for visual quality.

4.3. Method Comparison

Comparison with 2D-based Methods. In this section, we perform method comparisons with some representative 2D-based talking head generation approaches including the ATVGnet [5], MakeitTalk [54] and Wav2Lip [29]. Figure 7 visualizes the generated frames of these methods. It can be seen that the ATVGnet performs generation based on cropped faces with limited image quality. The MakeItTalk synthesizes plausible talking frames, however the background is wrongly wrapped with the mouth movements, which greatly affects the visual experience. Generated talking faces of Wav2Lip appear artifacts in the square boundary centered on the mouth, since the synthesized area and the original image are not well blended. By contrast, the proposed DiffTalk generates natural and realistic talking videos with accurate audio-lip synchronization, owing to the crafted conditioning mechanism and stable training process. For more objective comparisons, we further evaluate the quantitative results in Table 3. Our DiffTalk far surpasses [29] and [54] in all image quality metrics. For the audio-visual synchronization metric SyncNet, the proposed method reaches a high level and is superior to MakeItTalk. Although the DiffTalk is slightly inferior to Wav2Lip on SyncNet score, it is far better than Wav2Lip in terms of image quality. In conclusion, our method outperforms these

Method	Test Set A				Test Set B				General Method
	PSNR↑	SSIM↑	LPIPS↓	SyncNet↓↑	PSNR↑	SSIM↑	LPIPS↓	SyncNet↓↑	
GT	-	-	-	-1/8.979	-	-	-	-2/7.924	-
MakeItTalk [54]	18.77	0.544	0.19	-4/3.936	17.70	0.648	0.129	-3/3.416	✓
Wav2Lip [29]	25.50	0.761	0.140	**-2**/**8.936**	33.38	0.942	0.027	-3/**9.385**	✓
AD-NeRF [17]	27.89	0.885	0.072	**-2**/5.639	30.14	0.947	**0.023**	-3/4.246	×
DFRF [37]	**28.60**	**0.892**	**0.068**	**-1**/5.999	**33.57**	**0.949**	0.025	**-2**/4.432	FT Req.
Ours	**34.54**	**0.950**	**0.024**	**-1**/**6.381**	**34.01**	**0.950**	**0.020**	**-1**/**5.639**	✓

Table 3. Comparison with some representative talking head synthesis methods on two test sets as in Figure 7. The best performance is **highlighted** in **red** (1st best) and **blue** (2nd best). Our DiffTalk obtains the best PSNR, SSIM, and LPIPS values, and comparable SyncNet scores simultaneously. It is worth noting that the DFRF is fine-tuned on the specific identity to obtain these results, while our method can directly be utilized for generation without further fine-tuning. ('FT Req.' means that fine-tuning operation is required for the DFRF.)

(a) Resolution: 256 × 256, f=4 (b) Resolution: 512 × 512, f=8

Figure 8. Generated results with higher resolution.

2D-based methods under comprehensive consideration of the qualitative and quantitative results.

Comparison with 3D-based Methods. For more comprehensive evaluations, we further compare with some recent high-performance 3D-based works including AD-NeRF [17] and DFRF [37]. They realize implicitly 3D head modeling with the NeRF technology, so we treat them as generalized 3D-based methods. The visualization results are shown in Figure 7. AD-NeRF models the head and torso parts separately, resulting in misalignment in the neck region. More importantly, it is worth noting that AD-NeRF is a non-general method. In contrast, our method is able to handle unseen identities without further fine-tuning, which is more in line with the practical application scenarios. The DFRF relies heavily on the fine-tuning operation for model generalization, and the generated talking faces with only the base model are far from satisfactory as shown in Figure 7. More quantitative results in Table 3 also show that our method surpasses [17, 37] on the image quality and audio-visual synchronization indicators.

4.4. Expand to Higher Resolution

In this section, we perform experiments to demonstrate the capacity of our method on generating higher-resolution images. In Figure 8, we show the synthesis frames of two models (a) and (b). (a) is trained on 256×256 images with the downsampling factor $f = 4$, so the latent space is of size $64 \times 64 \times 3$. For (b), 512×512 images with $f = 8$ are used for training the model. Since both models are trained based on a compressed $64 \times 63 \times 3$ latent space, the pressure of insufficient computing resources is relieved. We can therefore comfortably expand our model for higher-resolution generation just as shown in Figure 8, where the synthesis quality in (b) significantly outperforms that in (a).

5. Conclusion and Discussion

In this paper, we have proposed a generalized and high-fidelity talking head synthesis method based on a crafted conditional diffusion model. Apart from the audio signal condition to drive the lip motions, we further incorporate reference images as driving factors to model the personalized appearance, which enables the learned model to comfortably generalize across different identities without any further fine-tuning. Furthermore, our proposed DiffTalk can be gracefully tailored for higher-resolution synthesis with negligible extra computational cost.

Limitations. The DiffTalk models talking head generation as an iterative denoising process, which needs more time to synthesize a frame compared with most GAN-based approaches. This is also a common problem of LDM-based works. When driving a portrait with more challenging cross-identity audio, the audio-lip synchronization of the synthesized video is slightly inferior to the ones under self-driven setting. During inference, the network is also sensitive to the mask shape in z_T, where the mask needs to cover the mouth region completely and its shape cannot leak any lip shape information. All these inspire our further research directions for superior synthesis results. Since talking head technology may raise potential misuse issues, we are committed to combating these malicious behaviors and advocate positive applications. Additionally, researchers who want to use our code will be required to get authorization and add watermarks to the generated videos.

Acknowledgment. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700802, in part by the National Natural Science Foundation of China under Grant 62125603, and in part by a grant from the Beijing Academy of Artificial Intelligence (BAAI).

References

[1] Miguel Angel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, et al. Gaudi: A neural architect for immersive 3d scene generation. *arXiv*, 2022. 3

[2] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999. 2

[3] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, 2021. 2

[4] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *ECCV*, 2020. 1, 2

[5] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *CVPR*, 2019. 7

[6] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: Video-and-audio-driven talking head synthesis. *arXiv*, 2020. 2

[7] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv*, 2022. 3

[8] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *ACCV*, 2016. 2, 5

[9] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *CVPR*, 2019. 4

[10] Dipanjan Das, Sandika Biswas, Sanjana Sinha, and Brojeshwar Bhowmick. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *ECCV*, 2020. 2

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021. 2

[12] Marcelo dos Santos, Rayson Laroca, Rafael O Ribeiro, João Neves, Hugo Proença, and David Menotti. Face super-resolution using stochastic differential equations. *arXiv*, 2022. 3

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2, 3

[14] Pablo Garrido, Levi Valgaerts, Hamid Sarmadi, Ingmar Steiner, Kiran Varanasi, Patrick Perez, and Christian Theobalt. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. In *Computer Graph. Forum*, 2015. 2

[15] Charles J Geyer. Practical markov chain monte carlo. *Statistical science*, 1992. 3

[16] Kuangxiao Gu, Yuqian Zhou, and Thomas Huang. Flnet: Landmark driven fetching and learning network for faithful talking facial animation synthesis. In *AAAI*, 2020. 2

[17] 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 *ICCV*, 2021. 2, 7, 8

[18] 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*, 2014. 4

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 2, 3

[20] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *ECCV*, 2022. 5

[21] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *CVPR*, 2021. 2

[22] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv*, 2022. 3

[23] Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and CV Jawahar. Towards automatic face-to-face translation. In *ACMMM*, 2019. 2

[24] Wing-Fung Ku, Wan-Chi Siu, Xi Cheng, and H Anthony Chan. Intelligent painter: Picture composition with resampling diffusion model. *arXiv*, 2022. 3

[25] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-aware implicit neural audio-driven video portrait generation. *arXiv*, 2022. 2

[26] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 3

[27] 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. In *ECCV*, 2020. 2

[28] Kushagra Pandey, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar. Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents. *arXiv*, 2022. 3

[29] 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 *ACMMM*, 2020. 2, 7, 8

[30] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 2

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 4, 5

[32] Robin Rombach, Andreas Blattmann, and Björn Ommer. Text-guided synthesis of artistic images with retrieval-augmented diffusion models. *arXiv*, 2022. 3

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2, 3

[34] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv*, 2022. 3

[35] Saeed Saadatnejad, Ali Rasekh, Mohammadreza Mofayezi, Yasamin Medghalchi, Sara Rajabzadeh, Taylor Mordan, and Alexandre Alahi. A generic diffusion-based approach for 3d human pose prediction in the wild. *arXiv*, 2022. 3

[36] Ruizhi Shao, Zerong Zheng, Hongwen Zhang, Jingxiang Sun, and Yebin Liu. Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. *arXiv*, 2022. 3

[37] 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 *ECCV*, 2022. 2, 7, 8

[38] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv*, 2020. 5

[40] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want. *arXiv*, 2020. 2

[41] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *TOG*, 2017. 2

[42] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *ECCV*, 2020. 2, 4

[43] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016. 2

[44] Dominik JE Waibel, Ernst Röoell, Bastian Rieck, Raja Giryes, and Carsten Marr. A diffusion model predicts 3d shapes from 2d microscopy images. *arXiv*, 2022. 3

[45] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, 2021. 1

[46] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 5

[47] Shunyu Yao, RuiZhe Zhong, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering. *arXiv*, 2022. 2

[48] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *ICCV*, 2019. 1

[49] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv*, 2022. 3

[50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2, 5, 7

[51] Xi Zhang, Xiaolin Wu, Xinliang Zhai, Xianye Ben, and Chengjie Tu. Davd-net: Deep audio-aided video decompression of talking heads. In *CVPR*, 2020. 1

[52] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *CVPR*, 2021. 5

[53] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *CVPR*, 2021. 2

[54] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *TOG*, 2020. 1, 2, 7, 8

[55] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. In *Computer Graphics Forum*, 2018. 2