

EMO: Emote Portrait Alive - Generating Expressive Portrait Videos with Audio2Video Diffusion Model under Weak Conditions

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo

Institute for Intelligent Computing, Alibaba Group
 {tianlinrui.tlr, wilson.wq, zhangbang.zb, liefeng.bo}@alibaba-inc.com
<https://humanaigc.github.io/emote-portrait-alive/>

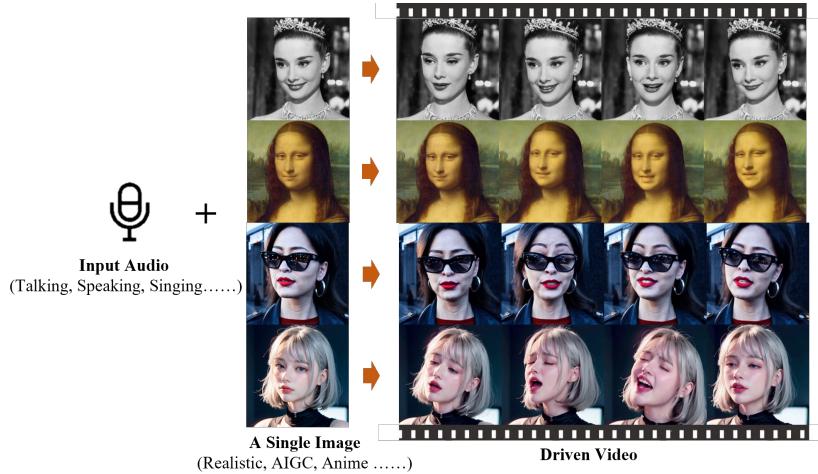


Fig. 1: We proposed EMO, an expressive audio-driven portrait-video generation framework. Input a **single reference image** and the vocal audio, e.g. talking and singing, our method can generate vocal avatar videos with expressive facial expressions, and various head poses, meanwhile, we can generate videos with **any duration** depending on the length of input audio.

Abstract. In this work, we tackle the challenge of enhancing the realism and expressiveness in talking head video generation by focusing on the dynamic and nuanced relationship between audio cues and facial movements. We identify the limitations of traditional techniques that often fail to capture the full spectrum of human expressions and the uniqueness of individual facial styles. To address these issues, we propose EMO, a novel framework that utilizes a direct audio-to-video synthesis approach, bypassing the need for intermediate 3D models or facial landmarks. Our method ensures seamless frame transitions and consistent

identity preservation throughout the video, resulting in highly expressive and lifelike animations. Experimental results demonstrate that EMO is able to produce not only convincing speaking videos but also singing videos in various styles, significantly outperforming existing state-of-the-art methodologies in terms of expressiveness and realism.

Keywords: Diffusion Models · Video Generation · Talking Head

1 Introduction

In recent years, the field of image generation has witnessed remarkable advancements, largely attributed to the emergence and success of Diffusion Models [4, 8, 16, 20, 25]. These models, celebrated for their ability to produce high-quality images, owe their prowess to extensive training on large-scale image datasets and a progressive generation approach. This innovative methodology enables the creation of images with unparalleled detail and realism, setting new benchmarks in the domain of generative models. The application of Diffusion Models has not been confined to still images alone. A burgeoning interest in video generation has led to the exploration of these models' potential in crafting dynamic and compelling visual narratives [6, 9]. These pioneering efforts underscore the vast potential of Diffusion Models in the field of video generation.

Beyond general video synthesis, the generation of human-centric videos has been the focal point of research, such as talking head. The objective of talking head is to generate the facial expressions from user-provided audio clips. Crafting these expressions involves capturing the subtleties and diversity of human facial movements, presenting a significant challenge in video synthesis. Traditional approaches often impose constraints on the final video output to simplify this task. For instance, some methods use 3D models to restrict facial keypoints, while others extract sequences of head movements from base videos to guide the overall motion. While these constraints reduce the complexity of video generation, they also tend to limit the richness and naturalness of the resulting facial expressions.

In this paper, our goal is to establish an innovative talking head framework designed to capture a broad spectrum of realistic facial expressions, including nuanced micro-expressions, and to facilitate natural head movements, thereby imbuing generated head videos with an unparalleled level of expressiveness. To achieve this goal, we propose a method that leverages the generative power of Diffusion Models, capable of directly synthesizing character head videos from a given image and an audio clip. This approach eliminates the need for intermediate representations or complex pre-processing, streamlining the creation of talking head videos that exhibit a high degree of visual and emotional fidelity, closely aligned with the nuances present in the audio input. Audio signals are rich in information related to facial expressions, theoretically enabling models to generate a diverse array of expressive facial movements. However, integrating audio with Diffusion Models is not a straightforward task due to the ambiguity inherent in the mapping between audio and facial expression. This issue can lead

to instability in the videos produced by the model, manifesting as facial distortions or jittering between video frames, and in severe cases, may even result in the complete collapse of the video. To address this challenge, we have incorporated stable control mechanisms into our model, namely a speed controller and a face region controller, to enhance stability during the generation process. These two controllers function as hyperparameters, acting as subtle control signals that do not compromise the diversity and expressiveness of the final generated videos. Furthermore, to ensure that the character in the generated video remains consistent with the input reference image, we adopted and enhanced the approach of ReferenceNet by designing a similar module, FrameEncoding, aimed at preserving the character’s identity across the video.

Finally, To train our model, we constructed a vast and diverse audio-video dataset, amassing over 250 hours of footage and more than 150 million images. This expansive dataset encompasses a wide range of content, including speeches, film and television clips, and singing performances, and covers multiple languages such as Chinese and English. The rich variety of speaking and singing videos ensures that our training material captures a broad spectrum of human expressions and vocal styles, providing a solid foundation for the development of EMO. We conducted extensive experiments and comparisons on the HDTF dataset, where our approach surpassed current state-of-the-art (SOTA) methods, including DreamTalk, Wav2Lip, and SadTalker, across multiple metrics such as FID, SyncNet, F-SIM, and FVD. In addition to quantitative assessments, we also carried out a comprehensive user study and qualitative evaluations, which revealed that our method is capable of generating highly natural and expressive talking and even singing videos, achieving the best results to date.

2 Related Work

Diffusion Models Diffusion Models have demonstrated remarkable capabilities across various domains, including image synthesis [4, 8], image editing [10, 24], video generation [6, 9] and even 3D content generation [12, 17]. Among them, Stable Diffusion (SD) [20] stands out as a representative example, employing a UNet architecture to iteratively generate images with notable text-to-image capabilities, following extensive training on large text-image datasets [23]. These pre-trained models have found widespread application in a variety of image and video generation tasks [6, 9]. Additionally, some recent works adopt a DiT (Diffusion-in-Transformer) [16] which alters the UNet with a Transformer incorporating temporal modules and 3D Convolutions, enabling support for larger-scale data and model parameters. By training the entire text-to-video model from scratch, it achieves superior video generation results [14]. Also, some efforts have delved into applying Diffusion Models for talking head generation, producing promising results that highlight the capability of these models in crafting realistic talking head videos [15, 27].

Audio-driven talking head generation Audio-driven talking head generation can be broadly categorized into two approaches: video-based methods

[5, 18, 30] and single-image [15, 28, 33]. video-based talking head generation allows for direct editing on an input video segment. For example, Wav2Lip [18] regenerates lip movements in a video based on audio, using a discriminator for audio-lip sync. Its limitation is relying on a base video, leading to fixed head movements and only generating mouth movements, which can limit realism. For single-image talking head generation, a reference photo is utilized to generate a video that mirrors the appearance of the photo. [28] proposes to generate the head motion and facial expressions independently by learning blendshapes and head poses. These are then used to create a 3D facial mesh, serving as an intermediate representation to guide the final video frame generation. Similarly, [33] employs a 3D Morphable Model (3DMM) as an intermediate representation for generating talking head video. A common issue with these methods is the limited representational capacity of the 3D mesh, which constrains the overall expressiveness and realism of the generated videos. Additionally, both methods are based on non-diffusion models, which further limits the performance of the generated results. [15] attempts to use diffusion models for talking head generation, but instead of applying directly to image frames, it employs them to generate coefficients for 3DMM. Compared to the previous two methods, Dreamtalk offers some improvement in the results, but it still falls short of achieving highly natural facial video generation.

3 Method

Given a single reference image of a character portrait, our approach can generate a video synchronized with an input voice audio clip, preserving the natural head motion and vivid expression in harmony with the tonal variances of the provided vocal audio. By creating a seamless series of cascaded video, our model facilitates the generation of long-duration talking portrait videos with consistent identity and coherent motion, which are crucial for realistic applications.

3.1 Preliminaries

Our methodology employs Stable Diffusion (SD) as the foundational framework. SD is a widely-utilized text-to-image (T2I) model that evolves from the Latent Diffusion Model (LDM) [20]. It utilizes an autoencoder Variational Autoencoder (VAE) [11] to map the original image feature distribution x_0 into latent space z_0 , encoding the image as $z_0 = \mathbf{E}(x_0)$ and reconstructing the latent features as $x_0 = \mathbf{D}(z_0)$. This architecture offers the advantage of reducing computational costs while maintaining high visual fidelity. Based on the Denoising Diffusion Probabilistic Model (DDPM) [8] or the Denoising Diffusion Implicit Model (DDIM) [26] approach, SD introduces Gaussian noise ϵ to the latent z_0 to produce a noisy latent z_t at a specific timestep t . During inference, SD aims to remove the noise ϵ from the latent z_t and incorporates text control to achieve the desired outcome by integrating text features. The training objective for this

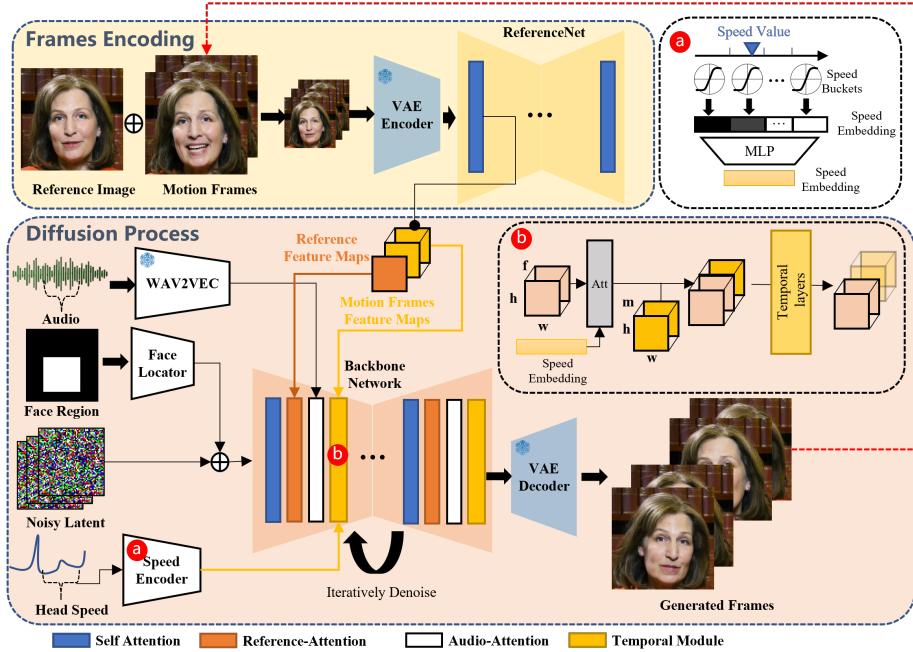


Fig. 2: Overview of the proposed method. Our framework is mainly constituted with two stages. In the initial stage, termed Frames Encoding, the ReferenceNet is deployed to extract features from the reference image and motion frames. Subsequently, during the Diffusion Process stage, a pretrained audio encoder processes the audio embedding. The facial region mask is integrated with multi-frame noise to govern the generation of facial imagery. This is followed by the employment of the Backbone Network to facilitate the denoising operation. Within the Backbone Network, two forms of attention mechanisms are applied: Reference-Attention and Audio-Attention. These mechanisms are essential for preserving the character’s identity and modulating the character’s movements, respectively. Additionally, Temporal Modules are utilized to manipulate the temporal dimension, and adjust the velocity of motion.

denoising process is expressed as:

$$\mathcal{L} = \mathbb{E}_{t,c,z_t,\epsilon} [||\epsilon - \epsilon_\theta(z_t, t, c)||^2] \quad (1)$$

where c represents the text features, which are obtained from the token prompt via the CLIP [19] ViT-L/14 text encoder. In SD, ϵ_θ is realized through a modified UNet [21] model, which employs the cross-attention mechanism to fuse c with the latent features.

3.2 Network Pipelines

The overview of our method is shown in Figure 2. Our **Backbone Network** get the multi-frame noise latent input, and try to denoise them to the consecutive video frames during each time step, the Backbone Network has the similar

UNet structure configuration with the original SD 1.5. 1) Similar to previous work, to ensure the continuity between generated frames, the Backbone Network is embedded with **temporal modules**. 2) To maintain the ID consistency of the portrait in the generated frames, we deploy a UNet structure called **ReferenceNet** parallel to the Backbone, it input the reference image to get the reference features. 3) To drive the character speaking motion, a **audio layers** is utilized to encode the voice features. 4) To make the motion of talking character controllable and stable, we use the **face locator** and **speed layers** to provide weak conditions.

Backbone Network. In our work, the prompt embedding is not utilized; hence, we have adapted the cross-attention layers in the SD 1.5 UNet structure to reference-attention layers. These modified layers now take reference features from ReferenceNet as input rather than text embeddings.

Audio Layers. The pronunciation and tone in the voice is the main driven sign to the motion of the generated character. The features extracted from the input audio sequence by the various blocks of the pretrained wav2vec [22] are concatenated to yield the audio representation embedding, $A^{(f)}$, for the f_{th} frame. However, the motion might be influenced by the future/past audio segments, for example, opening mouth and inhaling before speaking. To address that, we define voice features of each generated frame by concatenating the features of nearby frames: $A_{gen}^{(f)} = \oplus\{A^{(f-m)}, \dots, A^{(f)}, \dots, A^{(f+m)}\}$, m is the number of additional features from one side. To inject the voice features into the generation procedure, we add audio-attention layers performing a cross attention mechanism between the latent code and $A_{gen}^{(f)}$ after each ref-attention layers in the Backbone Network.

ReferenceNet. The ReferenceNet possesses a structure identical to that of the Backbone Network and serves to extract detailed features from input images. Given that both the ReferenceNet and the Backbone Network originate from the same original SD 1.5 UNet architecture, the feature maps generated by these two structures at certain layers are likely to exhibit similarities. Consequently, this facilitates the Backbone Network’s integration of features extracted by the ReferenceNet. Prior research [9, 35] has underscored the profound influence of utilizing analogous structures in maintaining the consistency of the target object’s identity. In our study, both the ReferenceNet and the Backbone Network inherit weights from the original SD UNet. The image of the target character is inputted into the ReferenceNet to extract the reference feature maps outputs from the self-attention layers. During the Backbone denoising procedure, the features of corresponding layers undergo a reference-attention layers with the extracted feature maps. As the ReferenceNet is primarily designed to handle individual images, it lacks the temporal layers found in the Backbone.

Temporal Modules. Most works try inserting temporal mixing layers into the pretrained text-to-image architecture, to facilitate the understanding and encoding of temporal relationships between consecutive video frames. By doing so, the enhanced model is able to maintain continuity and consistency across frames, resulting in the generation of smooth and coherent video streams. Informed by the architectural concepts of AnimateDiff, we apply self-attention temporal layers to the features within frames. Specifically, we reconfigure the input feature map $x \in \mathbb{R}^{b \times c \times f \times h \times w}$ to the shape $(b \times h \times w) \times f \times c$. Here, b stands for the batch size, h and w indicate the spatial dimensions of the feature map, f is the count of generated frames, and c is the feature dimension. Notably, we direct the self-attention across the temporal dimension f , to effectively capture the dynamic content of the video. The temporal layers are inserted at each resolution stratum of the Backbone Network. Most current diffusion-based video generation models are inherently limited by their design to produce a predetermined number of frames, thereby constraining the creation of extended video sequences. This limitation is particularly impactful in applications of talking head videos, where a sufficient duration is essential for the articulation of meaningful speaking. Some methodologies employ a frame from the end of the preceding clip as the initial frame of the subsequent generation, aiming to maintain a seamless transition across concatenated segments. Inspired by that, our approach incorporates the last n frames, termed 'motion frames' from the previously generated clip to enhance cross-clip consistency. Specifically, these n motion frames are fed into the ReferenceNet to pre-extract multi-resolution motion feature maps. During the denoising process within the Backbone Network, we merge the temporal layer inputs with the pre-extracted motion features of matching resolution along the frames dimension. This straightforward method effectively ensures coherence among various clips. For the generation of the first video clip, we initialize the motion frames as zero maps.

It should be noted that while the Backbone Network may be iterated multiple times to denoise the noisy frames, the target image and motion frames are concatenated and input into the ReferenceNet only once. Consequently, the extracted features are reused throughout the process, ensuring that there is no substantial increase in computational time during inference.

Face Locator and Speed Layers. Temporal modules can guarantee continuity of the generated frames and seamless transitions between video clips, however, they are insufficient to ensure the consistency and stability of the generated character's motion across the clips, due to the independent generation process. Previous works use some signal to control the character motion, e.g, skeleton [9], blendshape [33], or 3DMM [28], nevertheless, employing these control signals may be not good in creating lively facial expressions and actions due to their limited degrees of freedom, and the inadequate labeling during training stage are insufficient to capture the full range of facial dynamics. Additionally, the same control signals could result in discrepancies between different characters, failing to account for individual nuances. Enabling the generation of control

signals may be a viable approach [28], yet producing lifelike motion remains a challenge. Therefore, we opt for a "weak" control signal approach.

Specifically, as shown in Figure 2, we utilize a mask $M = \bigcup_{i=0}^f M^i$ as the face region, which encompasses the facial bounding box regions of the video clip. We employ the Face Locator, which consists of lightweight convolutional layers designed to encode the bounding box mask. The resulting encoded mask is then added to the noisy latent representation before being fed into the Backbone. In this way, we can use the mask to control where the character face should be generated.

However, creating consistent and smooth motion between clips is challenging due to variations in head motion frequency during separate generation processes. To address this issue, we incorporate the target head motion speed into the generation. More precisely, we consider the head rotation velocity w^f in frame f and divide the range of speeds into d discrete speed buckets, each representing a different velocity level. Each bucket has a central value c^d and a radius r^d . We retarget w^f to a vector $S = \{s^d\} \in \mathbb{R}^d$, where $s^d = \tanh((w^f - c^d)/r^d * 3)$. Similar to the method used in the audio layers, the head rotation speed embedding for each frame is given by $S^f = \oplus\{S^{(f-m)}, \dots, S^{(f)}, \dots, S^{(f+m)}\}$, and $S^f \in \mathbb{R}^{b \times f \times c^{speed}}$ is subsequently processed by an MLP to extract speed features. Within the temporal layers, we repeat S^f to the shape $(b \times h \times w) \times f \times c^{speed}$ and implement a cross-attention mechanism that operates between the speed features and the reshaped feature map across the temporal dimension f . By doing so and specifying a target speed, we can synchronize the rotation speed and frequency of the generated character's head across different clips. Combined with the facial position control provided by the Face Locator, the resulting output can be both stable and controllable.

It should also be noted that the specified face region and assigned speed does not constitute strong control conditions. In the context of face locator, since the M is the union area of the entire video clip, indicating a sizeable region within which facial movement is permissible, thereby ensuring that the head is not restricted to a static posture. With regard to the speed layers, the difficulty in accurately estimating human head rotation speed for dataset labeling means that the predicted speed sequence is inherently noisy. Consequently, the generated head motion can only approximate the designated speed level. This limitation motivates the design of our speed buckets framework.

3.3 Training Strategies

The training process is structured into three stages. The first stage is the image pretraining, where the Backbone Network, the ReferenceNet, and the Face Locator are taken into training, in this stage, the Backbone takes a single frame as input, while ReferenceNet handles a distinct, randomly chosen frame from the same video clip. Both the Backbone and the ReferenceNet initialize weights from the original SD. In the second stage, we introduce the video training, where the temporal modules and the audio layers are incorporated, $n+f$ contiguous frames are sampled from the video clip, with the started n frames are motion frames.

The temporal modules initialize weights from AnimateDiff [6]. In the last stage, the speed layers are integrated, we only train the temporal modules and the speed layers in this stage. This strategic decision deliberately omits the audio layers from the training process. Because the speaking character’s expression, mouth motion, and the frequency of the head movement, are predominantly influenced by the audio. Consequently, there appears to be a correlation between these elements, the model might be prompted to drive the character’s motion based on the speed signal rather than the audio. Our experimental results suggest that simultaneous training of both the speed and audio layers undermines the driven ability of the audio on the character’s motions.

4 Experiments

4.1 Implementations

We collected approximately 250 hours of talking head videos from the internet and supplemented this with the HDTF [34] and VFHQ [31] datasets to train our models. As the VFHQ dataset lacks audio, it is only used in the first training stage. We apply the MediaPipe face detection framework [13] to obtain the facial bounding box regions. Head rotation velocity was labeled by extracting the 6-DoF head pose for each frame using facial landmarks, followed by calculating the rotational degrees between successive frames.

The video clips sampled from the dataset are resized and cropped to 512×512 . In the first training stage, the reference image and the target frame are sampled from the video clip separately, we trained the Backbone Network and the ReferneceNet with a batch size of 48. In the second and the third stage, we set $f = 12$ as the generated video length, and the motion frames number is set to $n = 4$, we adopt a bath size of 4 for training. The additional features number m is set to 2, following the Diffused Heads [27]. The learning rate for all stages are set to 1e-5. During the inference, we use the sampling algorithm of DDIM to generate the video clip for 40 steps, we assign a constant speed value for each frame generation. The time consumption of our method is about 15 seconds for one batch ($f = 12$ frames).

4.2 Experiments Setup

For methods comparisons, we partitioned the HDTF dataset, allocating 10% as the test set and reserving the remaining 90% for training. We took precautions to ensure that there was no overlap of character IDs between these two subsets.

We compare our methods with some previous works including: Wav2Lip [18], SadTalker [33], DreamTalk [15]. Additionally, we generated results using the released code from Diffused Heads [27], however, the model is trained on CREMA [1] dataset which contains only green background, the generated results are suboptimal. Furthermore, the results were compromised by error accumulation across the generated frames. Therefore, we only conduct qualitative comparison

with the Diffused Heads approach. For DreamTalk, we utilize the talking style parameters as prescribed by the original authors.

To demonstrate the superiority of the proposed method, we evaluate the model with several quantitative metrics. We utilize Fréchet Inception Distance (FID) [7] to assess the quality of the generated frame [32]. Additionally, to gauge the preservation of identity in our results, we computed the facial similarity (F-SIM) by extracting and comparing facial features between the generated frames and the reference image. It is important to note that using a single, unvarying reference image could result in deceptively perfect F-SIM scores. Certain methods [18] might produce only the mouth regions, leaving the rest of the frame identical to the reference image, which could skew results. Therefore, we treat F-SIM as population-reference indices [27], with closer approximations to the corresponding ground truth (GT) values indicating better performance. We further employed the Fréchet Video Distance (FVD) [29] for the video-level evaluation. The SyncNet [2] score was used to assess the lip synchronization quality, a critical aspect for talking head applications. To evaluate the expressiveness of the facial expressions in the generated videos, we introduce the use of the Expression-FID (E-FID) metric. This involves extracting expression parameters via face reconstruction techniques, as described in [3]. Subsequently, we compute the FID of these expression parameters to quantitatively measure the divergence between the expressions in the synthesized videos and those in the ground truth dataset.

4.3 Qualitative Comparisons

Figure 3 demonstrates the visual results of our method alongside those of earlier approaches. It is observable that Wav2Lip typically synthesizes blurry mouth regions and produces videos characterized by a static head pose and minimal eye movement when a single reference image is provided as input. In the case of DreamTalk [15], the style clips supplied by the authors could distort the original faces, also constrain the facial expressions and the dynamism of head movements. In contrast to SadTalker and DreamTalk, our proposed method is capable of generating a greater range of head movements and more dynamic facial expressions. Since we do not utilize direct signal, e.g. blendshape or 3DMM, to control the character motion, these motions are directly driven by the audio, which will be discussed in detail in the following showcases.

We further explore the generation of talking head videos across various portrait styles. As illustrated in Figure 4, the reference images, sourced from Civitai, are synthesized by disparate text-to-image (T2I) models, encompassing characters of diverse styles, namely realistic, anime, and 3D. These characters are animated using identical vocal audio inputs, resulting in approximately uniform lip synchronization across the different styles. Although our model is trained only on the realistic videos, it demonstrates proficiency in producing talking head videos for a wide array of portrait types.

Figure 5 demonstrates that our method is capable of generating richer facial expressions and movements when processing audio with pronounced tonal



Fig. 3: The qualitative comparisons with several talking head generation works.

features. For instance, the examples in the third row reveal that high-pitched vocal tones elicit more intense and animated expressions from the characters. Moreover, leveraging motion frames allows for the extension of the generated video, we can generate prolonged duration video depending on the length of the input audio. As shown in Figure 5 and Figure 6, our approach preserves the character’s identity over extended sequences, even amidst substantial motion.

4.4 Quantitative Comparisons

Table 1: The quantitative comparisons with several talking head generation works.

Method	FID \downarrow	SyncNet \uparrow	F-SIM	FVD \downarrow	E-FID \downarrow
Wav2Lip [18]	9.38	5.79	80.34	407.93	0.693
SadTalker [33]	10.31	4.82	84.56	214.98	0.503
DreamTalk [15]	58.80	3.43	67.87	619.05	2.257
GT	-	7.3	77.44	-	-
Ours	8.76	3.89	78.96	67.66	0.116

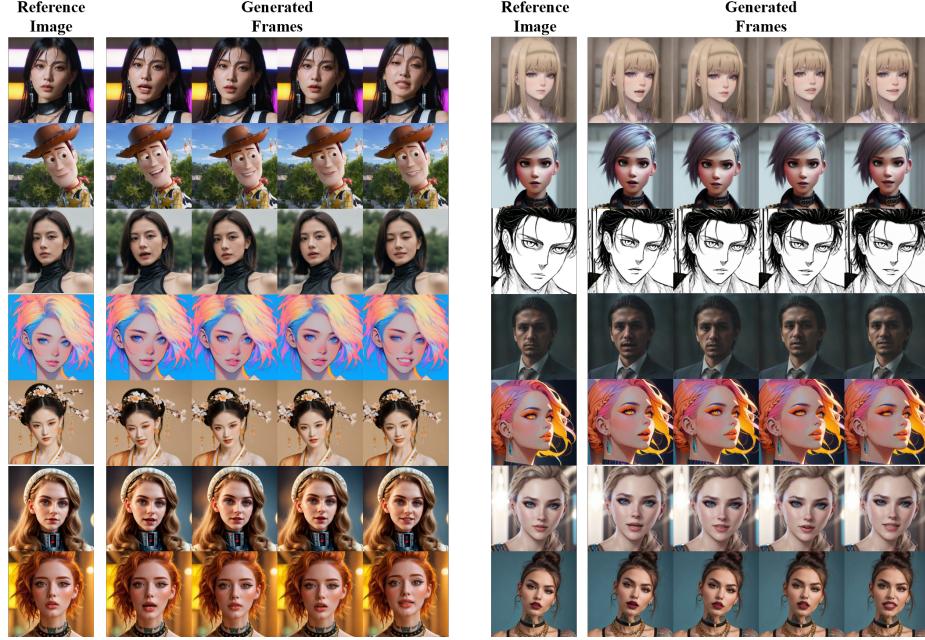


Fig. 4: The qualitative results of our method based on **different portrait styles**. Here we demonstrate 14 generated video clips, in which the characters are driven by the same vocal audio clip. The duration of each generated clip is approximately 8 seconds. Due to space limitations, we only sample four frames from each clip.

As shown in Table 1, our results demonstrate a substantial advantage in video quality assessment, as evidenced by the lower FVD scores. Additionally, our method outperforms others in terms of individual frame quality, as indicated by improved FID scores. Despite not achieving the highest scores on the SyncNet metric, our approach excels in generating lively facial expressions as shown by E-FID.

5 Limitation

There are some limitations for our method. First, it is more time-consuming compared to methods that do not rely on diffusion models. Second, since we do not use any explicit control signals to control the character’s motion, it may result in the inadvertent generation of other body parts, such as hands, leading to artifacts in the video. One potential solution to this issue is to employ control signals specifically for the body parts.

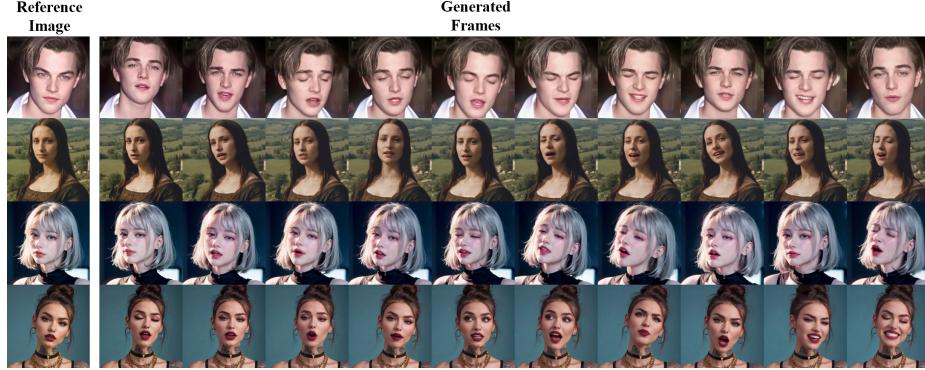


Fig. 5: The results generated by our method for vocal audio with a **strong tonal quality** over a **prolonged duration**. In each clip, the character is driven by the audio with strong tonal quality, e.g. singing, and the duration of each clip is approximately 1 minute.



Fig. 6: Comparisons with Diffused Heads [27], the duration of generated clips is 6 seconds, the results of Diffused Heads have low resolution and are compromised by error accumulation across the generated frames.

References

1. Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R.: Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing* **5**(4), 377–390 (2014) [9](#)
2. Chung, J.S., Zisserman, A.: 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**. pp. 251–263. Springer (2017) [10](#)
3. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: *IEEE Computer Vision and Pattern Recognition Workshops* (2019) [10](#)
4. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis (2021) [2](#), [3](#)
5. Fan, Y., Lin, Z., Saito, J., Wang, W., Komura, T.: Faceformer: Speech-driven 3d facial animation with transformers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022) [4](#)

6. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023) [2](#), [3](#), [9](#)
7. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017) [10](#)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) [2](#), [3](#), [4](#)
9. Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., Bo, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. arXiv preprint arXiv:2311.17117 (2023) [2](#), [3](#), [6](#), [7](#)
10. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mossner, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023) [3](#)
11. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. In: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings (2014) [4](#)
12. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Fidler, X.H.K.K.S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation [3](#)
13. Lugaressi, C., Tang, J., Nash, H., McClanahan, C., Ubowejia, E., Hays, M., Zhang, F., Chang, C.L., Yong, M., Lee, J., Chang, W.T., Hua, W., Georg, M., Grundmann, M.: Mediapipe: A framework for building perception pipelines (06 2019) [9](#)
14. Ma, X., Wang, Y., Jia, G., Chen, X., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048 (2024) [3](#)
15. Ma, Y., Zhang, S., Wang, J., Wang, X., Zhang, Y., Deng, Z.: Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. arXiv preprint arXiv:2312.09767 (2023) [3](#), [4](#), [9](#), [10](#), [11](#)
16. Peebles, W., Xie, S.: Scalable diffusion models with transformers. arXiv preprint arXiv:2212.09748 (2022) [2](#), [3](#)
17. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022) [3](#)
18. Prajwal, K.R., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: 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. p. 484–492. MM '20, Association for Computing Machinery, New York, NY, USA (2020). <https://doi.org/10.1145/3394171.3413532> [4](#), [9](#), [10](#), [11](#)
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) [5](#)
20. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) [2](#), [3](#), [4](#)
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing, Cham (2015) [5](#)

22. Schneider, S., Baevski, A., Collobert, R., Auli, M.: wav2vec: Unsupervised pre-training for speech recognition. pp. 3465–3469 (09 2019). <https://doi.org/10.21437/Interspeech.2019-1873> 6
23. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models (2022) 3
24. Shi, Y., Xue, C., Pan, J., Zhang, W., Tan, V.Y., Bai, S.: Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. arXiv preprint arXiv:2306.14435 (2023) 3
25. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015) 2
26. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021), <https://openreview.net/forum?id=St1giarCHLP> 4
27. Stypułkowski, M., Vougioukas, K., He, S., Zięba, M., Petridis, S., Pantic, M.: Diffused Heads: Diffusion Models Beat GANs on Talking-Face Generation. In: <https://arxiv.org/abs/2301.03396> (2023) 3, 9, 10, 13
28. Sun, X., Zhang, L., Zhu, H., Zhang, P., Zhang, B., Ji, X., Zhou, K., Gao, D., Bo, L., Cao, X.: Vividtalk: One-shot audio-driven talking head generation based on 3d hybrid prior. arXiv preprint arXiv:2312.01841 (2023) 4, 7, 8
29. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation (2019) 10
30. Wen, X., Wang, M., Richardt, C., Chen, Z.Y., Hu, S.M.: Photorealistic audio-driven video portraits. IEEE Transactions on Visualization and Computer Graphics **26**(12), 3457–3466 (2020). <https://doi.org/10.1109/TVCG.2020.3023573> 4
31. Xie, L., Wang, X., Zhang, H., Dong, C., Shan, Y.: Vfhq: A high-quality dataset and benchmark for video face super-resolution. In: The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2022) 9
32. Ye, Z., Zhong, T., Ren, Y., Yang, J., Li, W., Huang, J., Jiang, Z., He, J., Huang, R., Liu, J., et al.: Real3d-portrait: One-shot realistic 3d talking portrait synthesis. arXiv preprint arXiv:2401.08503 (2024) 10
33. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8652–8661. IEEE Computer Society, Los Alamitos, CA, USA (jun 2023). <https://doi.org/10.1109/CVPR52729.2023.00836>, <https://doi.ieee.org/10.1109/CVPR52729.2023.00836> 4, 7, 9, 11
34. Zhang, Z., Li, L., Ding, Y., Fan, C.: 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. pp. 3661–3670 (2021) 9
35. Zhu, L., Yang, D., Zhu, T., Reda, F., Chan, W., Saharia, C., Norouzi, M., Kemelmacher-Shlizerman, I.: Tryondiffusion: A tale of two unets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4606–4615 (2023) 6