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.

Sonic: Shifting Focus to Global Audio Perception in Portrait Animation

Xiaozhong Ji[1*] **Xiaobin Hu**[1*✉] **Zhihong Xu**[2] **Junwei Zhu**[1†] **Chuming Lin**[1] **Qingdong He**[1]
Jiangning Zhang[1] **Donghao Luo**[1] **Yi Chen**[1] **Qin Lin**[1] **Qinglin Lu**[1] **Chengjie Wang**[1]
[1] Tencent [2] Zhejiang University

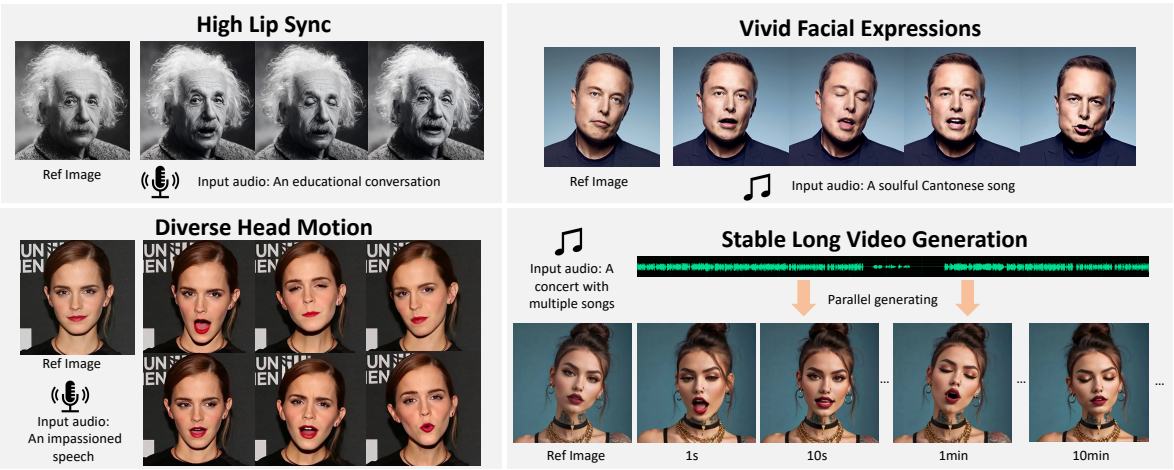

Figure 1. **Sonic excels in producing vivid portrait animation videos given a reference image and an audio clip.** Beyond its fundamental lip-syncing capabilities, Sonic demonstrates proficiency in creating a diverse spectrum of facial expressions and adaptable head movements. Notably, when dealing with extented videos, Sonic can yield stable and seamless outcomes in a parallel fashion, all while maintaining an unified paradigm focusing on global audio perception.

Abstract

The study of talking face generation mainly explores the intricacies of synchronizing facial movements and crafting visually appealing, temporally-coherent animations. However, due to the limited exploration of global audio perception, current approaches predominantly employ auxiliary visual and spatial knowledge to stabilize the movements, which often results in the deterioration of the naturalness and temporal inconsistencies. Considering the essence of audio-driven animation, the audio signal serves as the ideal and unique priors to adjust facial expressions and lip movements, without resorting to interference of any visual signals. Based on this motivation, we propose a novel paradigm, dubbed as Sonic, to shift focus on the exploration of global audio perception. To effectively leverage global audio knowledge, we disentangle it into intra- and inter-clip audio perception and collaborate with both aspects to enhance overall perception. For the intra-clip audio perception, 1). ***Context-enhanced audio learning****, in which long-range intra-clip temporal audio knowledge is extracted to provide facial expression and lip motion priors implicitly expressed as the tone and speed of speech. 2).* ***Motion-decoupled controller****, in which the motion of the head and expression movement are disentangled and independently controlled by intra-audio clips. Most importantly, for inter-clip audio perception, as a bridge to connect the intra-clips to achieve the global perception,* ***Time-aware position shift fusion****, in which the global inter-clip audio information is considered and fused for long-audio inference via through consecutively time-aware shifted windows. Extensive experiments demonstrate that the novel audio-driven paradigm outperform existing SOTA methodologies in terms of video quality, temporally consistency, lip synchronization precision, and motion diversity.*

1. Introduction

Talking Face Animation is a technique aimed at creating a lifelike representation of a speaking person by animating a still image in sync with speech audio. This technology holds great potential in enhancing the realism of virtual re-

* Equal Contribution, † Project Leader
✉ Corresponding Author
Project page: https://jixiaozhong.github.io/Sonic/.

ality [5], improving the efficiency of film and gaming making [25], facilitating accessibility for individuals with communication impairments [24], and offering therapeutic support and social interaction in healthcare [31]. Prominent progress in this domain is demonstrated by notable works based on Stable Diffusion [32], which project the parametric or implicit representations of lip movements [29, 50], facial expressions [16, 28, 33], and head movement [35] in the latent space [11, 27, 37, 45] to generate high-quality video with a high level of realism and liveliness.

However, precise audio control and temporal coherence have not yet been effectively achieved. Current approaches [8, 23, 37, 40, 42, 44, 49, 50] predominantly addressed these two issues in a completely segregated manner, overlooking the holistic coordination between audio and vision. For audio control, audio features were segmented according to timestamps and matched to each visual frame. By applying cross-attention within a spatial latent space, each frame was restricted to the nearby audio information, which hindered the transformation into an optimal motion representation. To achieve temporal coherence, self-attention along the temporal dimension was applied to smooth visual features within a clip, while overlapping or motion frames were used to enhance coherence between different clips. However, these stability strategies have limited temporal receptive fields, which can reduce the richness of motion. Additionally, they do not take audio information into account, potentially disrupting the synchronization between audio and visual elements.

To address these issues, we propose a novel paradigm, dubbed as Sonic, to extent the range of receptive field to the global level guided by the audio priors. To achieve global audio perception, we decompose the talking head video generation task into two aspects: *intra-clip* (*i.e.*, lip sync, expression alignment, and head movement) and *inter-clip* (*i.e.*, motion rhythm and smoothness, long video stability). Sonic aims to exploit global audio perception rather than the motion frames and other visual motion (*i.e.,* referring to the face locator and landmark prior adopted by existing methods, which hinders the ability to perceive information directly from audio). Performing as long-range and global signal, audio shows tone, rhythm and speed of speech which implicitly provides expression and head movement motion priors. However, we have to face such a tricky challenge that, as a weak correlation cross-modality signal without the guidance of any visual motion (*e.g.,* motion frames), current audio learning paradigm leads to the temporal jitter and unsatisfactory pattern and fails to well-align the audio and motion. To this end, we first customize efficient global audio perception learning mechanism. Specifically, **context-enhanced audio learning** is proposed as long-range audio learning module to extract the intra-clip audio temporal knowledge from the input audio clip. Then such intra-clip audio temporal information is mapped as a temporal-embeddings for the latter fusion by audio-temporal cross-attention. The temporal-embeddings projection operation can effectively reduce the computational burden and capture temporal priors within current audio clip. Although context-enhanced audio learning directs the temporal movements correlated with audio, such as lip movements synchronized with speech and facial expressions related to the content, there still exists the absence of habitual head movements. Secondly, we design **motion-decoupled controller** to disentangle the habitual head and expression movement and support the independent control via two explicit parameters learned from the current intra-clip audio. To enrich the functionality and playability of animation, we also add an plug to allow users to define some exaggerated movements. Most importantly, for inter-clip audio perception, **time-aware position shift fusion** enlarges the intra-clip audio perception to the global inter-clip audio perception via time-aware shifted windows consecutively bridging the preceding clip, providing the global inter-clip connection that significantly enhance modeling temporal power, as demonstrated in Figure 1. It is worth noting that our time-aware position shift fusion requires no extra training cost compared with the motion frames in most recent works, such as EMO [37], Loopy [23] and also no additional inference time caused by overlapping frames. Our contributions are summarized as:

- We propose a novel unified paradigm without the guidance of visual motion, dubbed as Sonic, to focus on the exploration of global audio perception.
- For intra-clip audio perception, context-enhanced audio learning and motion-decoupled controller excavate the intra-clip audio temporal knowledge to direct the temporal movements.
- For inter-clip audio perception, we propose time-aware position shift fusion to enhance the global inter-clip audio perception by extending the intra-clip audio perception through consecutively time-aware shifted windows.
- Extensive experiments demonstrate that the novel unified global audio perception paradigm outperforms existing SOTA methods in terms of video quality, temporal consistency, lip synchronization, and motion diversity.

2. Related Works

2.1. Diffusion-based Video Generation

The remarkable achievements in diffusion models in image generation have sparked widespread research interest for video generation. Early endeavors in video generation [14, 19, 20, 26, 34] have primarily concentrated on devising image-based diffusion models with temporal module to effectively capture the dynamic characteristics inherent in video sequences. To leverage the powerful abil-

ity of pretrained image diffusion models, Video LDM [4] incorporates temporal layers that learn to align images in a temporally consistent manner. AnimateDiff [15] provides a motion-specific module by multi-stage progressive training manner, and the motion module can be seamlessly integrated into existing text-to-image architectures without any further modifications. VideoComposer [41] and VideoCrafter [6] further explore the synthesis of image-to-video generation via textual and visual features combinations. Despite these notable advancements, the field still struggles to overcome the challenge of generating temporal coherent and high-quality videos. As a powerful motion representation trained on a large well-curated high-quality dataset, Stable Video Diffusion [3] shows the remarkable temporal-consistency and the high-quality visual characters for downstream tasks. There exists a large gap between the audio-driven portrait animation and general video generation due to the fact that audio-driven task aims to use the weak cross-modality audio signal to precisely control motion movements (*e.g.,* lip and expression movement).

2.2. Audio-driven Talking Face Generation

The generation of talking face videos from audio inputs has been a longstanding task. In the early stages, the emphasis was primarily on synthesizing lip movements alone, accomplished by directly mapping audio signals to lip movements while keeping other facial attributes unchanged [7, 22, 29, 36, 46]. In more recent efforts, the scope has been broadened to encompass a wider range of facial expressions and head movements, which are derived from audio inputs and a single reference image. These methods can be broadly classified into with or without intermediate representation. For instance, by employing 3D coefficients representation (*e.g.,* Sadtalker [50], Dreamtalk [28], AniPortrait [42], VASA-1 [45]), the video generation process can be informed through the direct or indirect visible control signal in the aspect of head movement and expression. Nevertheless, a recurring challenge faced by these techniques is the limited capacity of the 3D mesh to capture intricate details, consequently constraining the overall dynamism and realism. In contrast, the methods that omit intermediate representation exhibit superior naturalness and consistent preservation of identity with the original image. EMO [37], Hallo [44], and Loopy [23] employ a direct audio-to-video synthesis approach to ensure a high level of realism and naturalness, eliminating the need for intermediate 3D representation. However, these approaches primarily rely on additional conditions associated with spatial motion or motion frames to maintain temporal consistency. In this regard, they overlook the crucial role of audio in audio-driven animation, consequently diminishing both expressiveness and audio-motion synchronization.

2.3. Long Video Inference

Significant research efforts have been devoted to extending the duration of generated videos, aiming to expand their practical applications. Autoregressively predicting successive frames [17, 39] and hierarchical coarse-to-fine methods [47] have emerged to generate long videos. Then Lumiere [2] divides the video into overlapping temporal segments, denoises each segment independently, and finally fuse these segments in an optimization algorithm. However there is no available high-efficiency method for audio-condition global perception. As one step towards achieving stronger capabilities, our Sonic proposes time-aware position shift fusion strategy progressively denoise intra-clip audio to establish a global inter-clip connection through the use of time-aware shifted windows that bridge the preceding clip along the timesteps axis. This inter-clip fusion with global receptive field needs no additional training cost and introduce no extra inference time without overlapping frames.

3. Methodology

Overall framework. Given a single portrait reference image and an input audio, our one-stage framework can generate a portrait video well-synchronized with the driven audio. Our proposed Sonic framework aims to focus on the exploration of global audio perception, primarily consisting of context-enhanced audio learning, motion-decoupled controller, and time-aware position shift fusion module, as illustrated in Figure 2 .

Technical contribution. To achieve global audio perception, we decompose the talking head video generation task into two aspects: *intra-clip* (*i.e.*, lip sync, expression alignment, and head movement) and *inter-clip* (*i.e.*, motion rhythm and smoothness, long video stability). For intra-clip, we enhance the visual realism that matches audio signal through the context-enhanced audio learning and the motion controller. For inter-clip, we improve the stability using the time-aware position shift fusion strategy.

3.1. Context-enhanced Audio Learning

For audio-talking face animation, the audio is the ideally unique and global signal to adjust the face expression and movement without interference from other visual signals. Thus, the motivation behind context-enhanced audio learning is to guide the generation of spatial and temporal frames solely based on audio signals, rather than motion frames or other visual signals, to achieve the high-realism generation. Specifically, we adopt Whisper-Tiny [30] as the pretrained audio feature extraction model, which is more lightweight than Wav2Vec [1] used by most existing methods [8, 37, 44]. For each audio frame, the features from the last layers of five stages are concatenated together for multi-scale understanding. For each video frame, a duration of

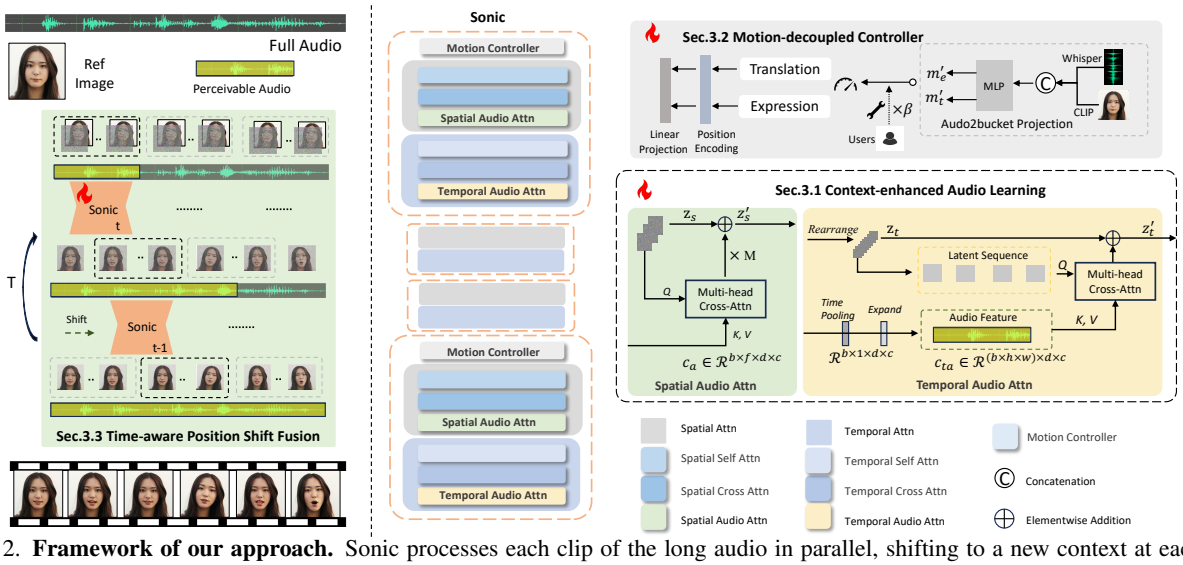

Figure 2. **Framework of our approach.** Sonic processes each clip of the long audio in parallel, shifting to a new context at each time step to progressively fuse inter-clip latent features across global audio perception. In Sonic, we enhance intra-clip temporal audio context learning and decouple motion to improve dynamics.

$0.2s$ audio feature is used frame-wise to give rich context. We use three linear layers to project audio feature to match the dimension of cross attention. The transformed audio embedding is denoted as $c_a \in \mathcal{R}^{b\times f\times d\times c}$, where b denotes batch size, f is audio length, d is num of context tokens, and c is hidden size of the cross-attention. To focus on talking face area, the core latent features is restricted with a mask M determined by a joint bounding boxes of faces. The audio signal can act as the spatial knowledge provider being injected into the spatial cross-attention layers as follows:

$$z_s^{'} = z_s + \mathrm{CrossAttn}(Q(z_s), K(c_a), V(c_a)) \cdot M, \tag{1}$$

where z_s is spatial latent features of self-attention layer, and $z_s^{'}$ is the adjusted spatial features guided by audio signals in spatial-aware level.

Furthermore, audio is a long-range and global signal that provides expression and head movement motion priors implicitly expressed through the tone and speed of speech. To excavate the temporal information from the long-range audio knowledge, we customize a temporal-audio module for precise motion control. The temporal module in existing methods [37, 50] follows AnimateDiff, which performs only self-attention across the temporal dimension without audio aids. Some other talking-face [23, 44] works extract the motion priors from the visual knowledge (*e.g.,* motion frames), which deviates from the essence of audio-driven task that relies solely on audio to drive movement. Thus we propose a temporal audio cross-attention module, introducing audio embedding to guide temporal alignment.

The preceding latent from temporal self-attention layers can be reshaped as $z_t \in \mathcal{R}^{(b\times h\times w)\times f\times c}$. To reduce the computation burden, the audio feature c_a is average-pooled along temporal dimension as the shape $\mathcal{R}^{(b\times 1)\times d\times c}$, then repeated and reshaped as $c_{ta} \in \mathcal{R}^{(b\times h\times w)\times d\times c}$. The projected audio temporal-embeddings are injected into the denoising U-Net via a temporal cross-attention blocks as:

$$z_t^{'} = z_t + \mathrm{CrossAttn}(Q(z_t), K(c_{ta}), V(c_{ta})). \tag{2}$$

The spatial and temporal audio attention layers are cascaded in multiple down-stages and up-stages to enable natural appearance and smooth temporal movement.

3.2. Motion-decoupled Controller

Context-enhanced Audio Learning governs the temporal movements strongly correlated with audio, such as lip movements synchronized with speech and facial expressions related to the content. On the other hand, our motion-decoupled controller performs a complementary role by directing the temporal movements weakly associated with audio, such as habitual head movements and random changes in perspective. This controller can disentangle motion into explicit head and expression movement amplitudes, thereby enhancing both functionality, playability, and interactivity. Thus, we introduce independent augmented motion-bucket parameters which effectively influence motion patterns of head movements and expression strength. During the training phase, the translation motion-bucket m_t is computed as the variance of bounding boxes of video clip, while the expression motion-bucket m_e is variance of relative landmarks. Both buckets are integers in the range of [0, 128]. Through position encoding and linear projection, they are added into ResNet blocks as follows:

$$emb = W[\mathrm{Pe}(m_t), \mathrm{Pe}(m_e)], \tag{3}$$

where emb is the embedding, Pe is the position encoding, and W represents the linear projection weights.

For practice application, motion-decoupled controller also supports the automatically adjustable parameters learning from the audio and reference image, and deeply explore the correlation between audio and portrait reference without requiring the parameter setting from users. To simplify multiple parameters and maintain adjustability, the predicted motion buckets are multiplying by a scale β (0.5 for mild dynamic, 1.0 for moderate dynamic, and 2.0 for intense dynamic). Specifically, we utilize a audio-to-bucket manner to adaptively predict the motion-buckets relying on the audio and reference image. In inference phase, the motion buckets are predicted and rescaled as:

$$m_t^{'}, m_e^{'} = \beta \times \mathrm{E_b}(c_a, R_{img}). \tag{4}$$

where $m_t^{'}$, $m_e^{'}$ are the estimated translation and expression motion-bucket achieved by the audio c_a and single reference image CLIP embedding R_{img}, and $\mathrm{E_b}$ is the audio-to-bucket projection consisting of there linear layers with the ReLU as activation function.

3.3. Time-aware Position Shift Fusion

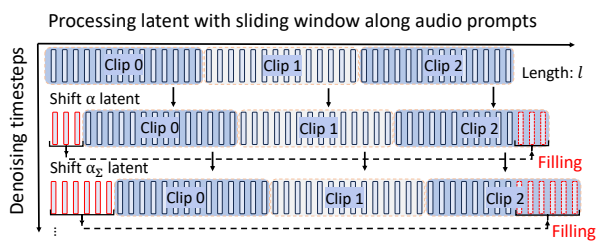

Figure 3. **Illustration of the proposed time-aware position-shift fusion.** The model processes each clip non-overlapping. In next timestep, the model starts from a new position determined by the offset, thereby integrating long-range context. Specifically, the tail latents are filled cyclically from the head.

Recent methods [23, 37, 44] use motion frames or overlapping latent during coherent denoising, which suffer from a limited receptive field and increased the training or inference computational complexity for audio-conditioned video generation problem. In contrast, we propose a bridge, the time-aware position shift fusion, to extent the intra-clip audio perception to the global inter-clip audio-perception via consecutively time-aware shifted windows. The simple yet effective fusion strategy, without increasing extra inference or training time costs, can effectively address jitter and abrupt transition issues. Audio serves as the input condition for the talking face, where longer audio prompts the generation of longer videos. Based on this principle, we tailor time-aware position-shift fusion mechanism along the timestep axis where the latent sliding window is progressively shifted by a certain length at different timesteps while also preserving the frame correspondence relationship between audio and video signal.

Specifically, as described in Figure 3 and Algorithm 1, our time-aware position-shift fusion consist of two nested loops. The outer loop is inverse diffusion process, while the inner loop is a sliding window process that model predicts for the full-length audio conditions. At each timestep, the model takes a clip of audio input to predict the corresponding latent with a start offset to smooth the segment from the last timestep. The video diffusion model naturally bridges the context, generating continuous movement following the audio prompts. In detail, we set a cumulative shift offset α_Σ, accumulating α at each timestep to ensure that the model starts sliding window from different positions. The offset α is experimentally set to a small integer, such as $3, 7$; we will report analysis on this parameters in experiments part. Note that the relative positioning between the latents and the audio prompts has not been altered. The only modification occurs in their recombination during the input phase to the model. In a special case, at the end of the long sequence, the shifted position index may exceed the sequence length. To address this issue, we adopt a circular padding strategy, padding the start audio prompt and latents to the end.

Algorithm 1 Time-aware Position-shift Fusion

Require: Audio embedding $c_a^{[0,l]}$ with length l, denoising steps T, initial noisy latent $z_T^{[0,l]}$, pretrained Sonic model $\mathrm{ST}(\cdot)$ for sequence length f, position-shift offset $\alpha < f < l$.
Ensure: Denoised latent $z_0^{[0,l]}$.

```
Initialize accumulated shift offset α_Σ = 0.
for t = T, ··· , 1 do
    // Denoising loop
    Initialize start point s = α_Σ, end e = s + f, processed length n = 0. // start from new position for each timestep.
    while n < l do
        // Sequence loop
        z_{t-1}^{[s,e]} = ST(z_t^{[s,e]}, c_a^{[s,e]}, t)
        s ← s + f, e ← e + f, n ← n + f. // Move to next clip non-overlapping
        if s > l or e > l then
            s ← s%l, e ← e%l. // Padding circularly
        end if
    end while
    α_Σ ← α_Σ + α. // Accumulate shift offset
end for
return Denoised latent z_0^{[0,l]}.
```

Computation cost. We analyze the computation efficiency of our time-aware position shift fusion compared to motion frames and overlapping. Denote the FLOPs of a single forward pass of the video diffusion model as Ω, the number of frames to generate as $n \times f$, and the denoising steps as T. The accumulated FLOPs of Sonic is: $\Omega(Shift) = \Omega \times T \times n$. For overlapping, the overlapped latents (overlap o) are calculated twice, $\Omega(Overlap) = \Omega \times T \times (n + \frac{o}{f})$. For motion frames, the FLOPs of a single forward of Ref-

Table 1. **Quantitative comparisons with the state-of-the-arts on the HDTF and CelebV-HQ test set.** * indicates GAN-based methods and others are Diffusion-based methods. The best results are **bold**, and the second are underlined.

Dataset	Method	FID↓	FVD↓	Sync-C↑	Sync-D↓	E-FID↓	F-SIM↑	Smooth↑	Runtime↓
HDTF	SadTalker* (2023) [50]	61.672	397.114	1.755	10.695	2.482	0.9287	0.9961	3.75
	Aniportrait (2024) [42]	36.965	471.452	1.095	12.461	3.161	0.9508	0.9921	44.03
	V-Express (2024) [40]	47.396	758.023	1.256	12.394	2.634	0.9093	0.9968	39.04
	Hallo (2024) [44]	30.176	347.358	4.060	9.551	1.792	0.9555	0.9941	74.65
	Hallo2 (2024) [12]	38.673	328.540	4.136	9.465	2.203	**0.9606**	0.9942	45.75
	EchoMimic (2024) [8]	33.207	384.304	2.514	10.743	**1.486**	0.9527	0.9934	5.45
	Sonic (Ours)	**29.104**	**301.173**	**4.197**	**9.371**	1.745	0.9595	**0.9970**	17.04
CelebV-HQ	SadTalker* (2024) [50]	57.574	841.962	1.978	10.915	2.252	0.9434	0.9959	3.75
	Aniportrait (2024) [42]	53.746	590.373	0.996	12.084	3.296	0.9522	0.9911	44.03
	V-Express (2024) [40]	65.400	889.985	0.809	13.255	2.713	0.9019	0.9965	39.04
	Hallo (2024) [44]	47.403	488.499	2.680	10.292	2.273	0.9607	0.9942	74.65
	Hallo2 (2024) [12]	52.396	**481.336**	2.638	10.343	2.819	0.9587	0.9942	45.75
	EchoMimic (2024) [8]	48.267	596.870	1.949	10.754	2.136	0.9612	0.9932	5.45
	Sonic (Ours)	**43.137**	483.108	**2.689**	**10.194**	**1.783**	**0.9624**	**0.9972**	17.04

Figure 4. **Qualitative comparisons with State-of-the-Art talking head generation methods.** Due to image does not reflect important sync, naturalness and stability, the full video comparison will be included in supplementary materials as well as comparison with demos from other non-open source works.

erence Net is ω_r, the number of motion frames is o and the Flops of motion modules computing f frames is ω_m, $\Omega(MF) = \Omega \times T \times n + \omega_r \times o \times n + \omega_m \times T \times n \times \frac{2of+o^2}{f^2}$. Overall, the proposed time-aware position shift fusion significantly reduces the computation complexity.

4. Experiments

4.1. Experiments Setup

Implementation Details. The Sonic training is in a single stage, with initialization of the spatial module and temporal modules from stable-video-diffusion-xt-1-1[3]. To separately enable different conditions in training, we manipulate the data such that 5% of it drops audio, another 5% drops image, and 5% drops both condition. In inference phase, multi-condition CFG is performed. The CFG of the reference image r_i is set to 2.0, and the CFG of the audio r_a is set to 7.5. In our full model, the dynamic scale β is set to 1.0 and the shift offset α is set to 7.

Datasets and Evaluation Metrics. For the training set, we created a diverse dataset by aggregating data from open-source datasets, including VFHQ [43], CelebV-Text [48], VoxCeleb2 [10]. For the test set, we conducted experiments on two widely used datasets, HDTF [52] and CelebV-HQ [51]. We randomly selected 50 four-second clips from CelebV-HQ and 100 from HDTF. We utilize seven met-

Figure 5. **Qualitative results under different styles of portrait images and various types of audio inputs.** The images and audios were collected from recent works and the Internet. The upper section presents results for audio input ranging from 20 seconds to 2 minutes in duration, while the lower section shows results for longer audio inputs, up to 10 minutes. Our Sonic demonstrates versatility across various portrait styles and maintains vividness over extended durations.

rics to evaluate. The Fréchet Inception Distance (FID) [18] and Fréchet Video Distance (FVD) [38] metrics measure the quality of the generated data. Expression-FID (E-FID) [37] uses face reconstruction method [13] to extract features and calculate the FID between generated and ground truth frames, while F-SIM measures facial similarity. SyncNet (Sync-C&Sync-D) [9] and VBench's smooth metric [21] assess lip synchronization and motion fluidity. In ablation experiments, we employ VBench's dynamic degree metric to evaluate amplitude of head movements.

4.2. Results and Analysis

Quantitative Comparison. In Table 1, our Sonic are significantly superior than other recent 6 methods in terms of FID, Sync-confidence and Smoothness evaluated on the CelebV-HQ and HDTF datasets. On the HDTF dataset, our Sonic achieve the lowest FID and FVD scores, 8% lower than the second-best, indicating superior generation quality and alignment with real data. For the CelebV-HQ dataset, which features more complex characters and backgrounds, our method maintains superior performance across most metrics and competitive FVD. We notably excel in E-FID, 16.5% lower than the second-best, which demonstrates our method has superior expression diversity. EMO [37] as a representative work that tackles talking head generation task, remains closed source, and not available for fair quantitative comparison. We provide video results using input from their online demos for a qualitative comparison.

Qualitative Comparison. Figure 4 provides a visualization comparison on open datasets. Analyzing the results, Previous approaches relied on motion frames, limiting the richness of generated expressions and the dynamism of head movements. In contrast, our proposed framework incorporates global audio perception, enhancing the integration of audio and temporal variables while providing a deeper understanding of the audio input. This enables our method to generate a broader range of expressions that align with the audio and facilitate more natural head movements.

Visualization Results in Complex Scenarios. We further explore the generation performance of our method under different input conditions. In terms of images, we employ three styles: real human, animation, and 3D. Each character portrait is paired with various types of audio relevant to its theme, such as singing, speech, and rap. As shown in Figure 5, our method successfully generates a wide array of expressions and diverse movements across various categories of input conditions, demonstrating strong robustness.

Table 2. User study comparison on open dataset.

Method / Metric	Lip sync	Motion diversity	ID consistency	Video Smoothness
Aniportrait	1.42	1.62	3.11	2.09
SadTalker	1.98	2.34	2.95	2.95
Echomimic	2.77	2.65	3.48	2.71
Hallo2	3.15	2.37	3.34	2.94
Sonic(Ours)	**4.58 (45%↑)**	**4.55 (72%↑)**	**4.29 (23%↑)**	**4.66 (58%↑)**

User Studies We conducted subjective evaluation on open dataset to assess the comparative methods along four key dimensions: Lip sync, motion diversity, identity similarity, and video smoothness. Total 40 participants gives scores from 1 to 5 for results of five comparative methods. In Table 2, our Sonic outperforms the others in all four dimensions, especially showing a significant 72% improvement in motion diversity. The video results for user study will be included in the supplementary materials.

Figure 6. Visual results for ablation on temporal audio attention and motion controller.

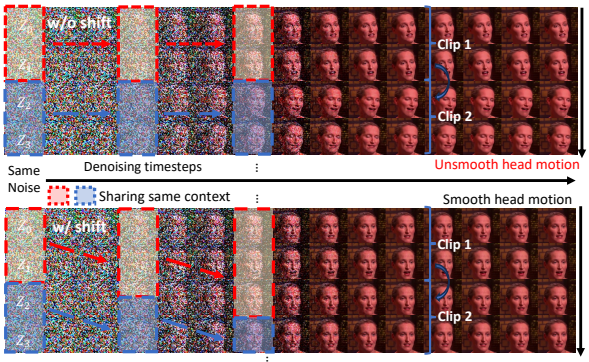

Figure 7. Demonstration on effect of time-aware position shift fusion.

Table 3. Quantitative results of ablations on CelebV-HQ along with our results obtained using different dynamic scales. In the Sonic model, the default setting for β is 1.0, *i.e.* Ours-Moderate.

Method / Metric	Sync-C↑	Sync-D↓	E-FID↓	Smooth↑	Dynamic↑
w/o temporal-audio	2.610	10.310	2.311	0.9969	0.62
w/o motion-controller	2.193	10.731	3.231	0.9969	0.22
w/o shift-fusion	2.509	10.379	2.202	0.9969	0.76
Ours-Mild ($\beta = 0.5$)	2.655	10.242	2.060	0.9970	0.78
Ours-Intense ($\beta = 2.0$)	**2.803**	**10.125**	2.038	0.9969	**0.98**
Ours-Moderate ($\beta = 1.0$)	2.689	10.194	**1.783**	**0.9972**	0.86

Table 4. Quantitative comparison on motion frames, overlap and the proposed time-aware shift fusion.

Method / Metric	Sync-C↑	Sync-D↓	Smooth↑	Runtime↓
Motion frames(8)	2.465	10.323	0.9969	19.70
Overlap(8)	2.544	10.265	0.9969	22.49
Shift-fusion	**2.689**	**10.194**	**0.9972**	**17.04**

4.3. Ablation Studies

Temporal Control Inspires Intra-clip Diversity. In Table 3, the absence of temporal audio attention results in a decrease in both lip-sync and E-FID. We observe that the motion controller plays a crucial role in CelebV-HQ given that its videos exhibit more movement that weakly correlates with audio. We also investigate the effect of different dynamic scale. Increasing β can further stimulate the lip-audio synchronization and the diversity of movements, while there is a slight decrease in stability and expression similarity. The instances in Figure 6 indicates temporal audio attention helps learn effective control of micro-expressions, while the motion controller is critical for agility and expressiveness.

Figure 8. Visual results compared with motion frames and overlapping.

Larger Audio Perception Bridges Inter-clip Stability. Initially, we ablate the time-aware position shift fusion (briefly termed Shift) and generate clips independently as illustrated in Figure 7. The fusion process is performed at each timestep, gradually inverting the random noise to continuous and smooth long range videos. Subsequently, we explore different long video inference strategy: motion frames, overlap, our shift. The number of motion frames and overlap frames was set to 8. As indicated in Table 4, our shift approach attains superior lip-sync and smoothness without necessitating any auxiliary computation. From the qualitative comparison demonstrated in Figure 8, our method successfully handle long term consistency, circumventing identity similarity decrease in motion frames and mixed blurry texture in overlapping results.

5. Conclusion

In summary, our work present Sonic, an audio-driven portrait animation that focus on the exploration of global audio perception to enhance the efficient generation of realistic lip synchronization, naturalness and temporal consistencies. The technical cornerstone mainly lies on context-enhanced audio learning and time-aware position shift fusion that work together to extent the intra-clip audio temporal to the global inter-clip audio perception. It significantly outperforms existing SOTA methods in video quality, motion diversity and naturalness.

Acknowledgements We thank Dongshan Li, Zixiang Zhou, Zhentao Yu, Hongmei Wang, Joeyu Wang, Chengfei Cai, Shiyu Tang, Tianxiang Zheng, Junshu Tang, Shuai Shao, Yubin Zeng, Junxin Huang, Yue Zhang, Chao Zhan, Zhengyan Tong, Zhaokang Chen, Bin Wu, Wenjiang Zhou, Yuang Zhang, Junqi Cheng, Jiaxi Gu, and Fangyuan Zou for helpful discussion.

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 3

[2] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024. 3

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3, 6

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 3

[5] Aras Bozkurt, Xiao Junhong, Sarah Lambert, Angelica Pazurek, Helen Crompton, Suzan Koseoglu, Robert Farrow, Melissa Bond, Chrissi Nerantzi, Sarah Honeychurch, et al. Speculative futures on chatgpt and generative artificial intelligence (ai): A collective reflection from the educational landscape. *Asian Journal of Distance Education*, 18(1):53–130, 2023. 2

[6] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. 3

[7] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European conference on computer vision (ECCV)*, pages 520–535, 2018. 3

[8] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024. 2, 3, 6

[9] Joon Son Chung and Andrew Zisserman. 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*, pages 251–263. Springer, 2017. 7

[10] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 6

[11] Enric Corona, Andrei Zanfir, Eduard Gabriel Bazavan, Nikos Kolotouros, Thiemo Alldieck, and Cristian Sminchisescu. Vlogger: Multimodal diffusion for embodied avatar synthesis. *arXiv preprint arXiv:2403.08764*, 2024. 2

[12] Jiahao Cui, Hui Li, Yao Yao, Hao Zhu, Hanlin Shang, Kaihui Cheng, Hang Zhou, Siyu Zhu, and Jingdong Wang. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. *arXiv preprint arXiv:2410.07718*, 2024. 6

[13] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 7

[14] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 2

[15] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024. 3

[16] Tianyu He, Junliang Guo, Runyi Yu, Yuchi Wang, Jialiang Zhu, Kaikai An, Leyi Li, Xu Tan, Chunyu Wang, Han Hu, et al. Gaia: Zero-shot talking avatar generation. *arXiv preprint arXiv:2311.15230*, 2023. 2

[17] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 3

[18] 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. *Advances in neural information processing systems*, 30, 2017. 7

[19] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[20] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2

[21] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 7

[22] Xiaozhong Ji, Chuming Lin, Zhonggan Ding, Ying Tai, Jian Yang, Junwei Zhu, Xiaobin Hu, Jiangning Zhang, Donghao Luo, and Chengjie Wang. Realtalk: Real-time and realistic audio-driven face generation with 3d facial prior-guided identity alignment network. *arXiv preprint arXiv:2406.18284*, 2024. 3

[23] Jianwen Jiang, Chao Liang, Jiaqi Yang, Gaojie Lin, Tianyun Zhong, and Yanbo Zheng. Loopy: Taming audio-driven portrait avatar with long-term motion dependency. *arXiv preprint arXiv:2409.02634*, 2024. 2, 3, 4, 5

[24] Esperanza Johnson, Ramón Hervás, Carlos Gutiérrez López de la Franca, Tania Mondéjar, Sergio F Ochoa, and Jesús

Favela. Assessing empathy and managing emotions through interactions with an affective avatar. *Health informatics journal*, 24(2):182–193, 2018. 2

[25] Oytun Kal and Yavuz Samur. Educational virtual reality game design for film and animation. *Encyclopedia of Computer Graphics and Games*, pages 621–636, 2024. 2

[26] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023. 2

[27] Tao Liu, Feilong Chen, Shuai Fan, Chenpeng Du, Qi Chen, Xie Chen, and Kai Yu. Anitalker: Animate vivid and diverse talking faces through identity-decoupled facial motion encoding. *arXiv preprint arXiv:2405.03121*, 2024. 2

[28] Yifeng Ma, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yingya Zhang, and Zhidong Deng. Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv preprint arXiv:2312.09767*, 2023. 2, 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 *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492, 2020. 2, 3

[30] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 3

[31] Imogen C Rehm, Emily Foenander, Klaire Wallace, Jo-Anne M Abbott, Michael Kyrios, and Neil Thomas. What role can avatars play in e-mental health interventions? exploring new models of client–therapist interaction. *Frontiers in Psychiatry*, 7:186, 2016. 2

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[33] Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1982–1991, 2023. 2

[34] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[35] Xusen Sun, Longhao Zhang, Hao Zhu, Peng Zhang, Bang Zhang, Xinya Ji, Kangneng Zhou, Daiheng Gao, Liefeng Bo, and Xun Cao. Vividtalk: One-shot audio-driven talking head generation based on 3d hybrid prior. *arXiv preprint arXiv:2312.01841*, 2023. 2

[36] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 3

[37] 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. 2, 3, 4, 5, 7

[38] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 7

[39] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in neural information processing systems*, 35:23371–23385, 2022. 3

[40] Cong Wang, Kuan Tian, Jun Zhang, Yonghang Guan, Feng Luo, Fei Shen, Zhiwei Jiang, Qing Gu, Xiao Han, and Wei Yang. V-express: Conditional dropout for progressive training of portrait video generation. *arXiv preprint arXiv:2406.02511*, 2024. 2, 6

[41] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[42] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024. 2, 3, 6

[43] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 657–666, 2022. 6

[44] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Luc Van Gool, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024. 2, 3, 4, 5, 6

[45] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *arXiv preprint arXiv:2404.10667*, 2024. 2, 3

[46] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European conference on computer vision*, pages 85–101. Springer, 2022. 3

[47] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, et al. Nuwa-xl: Diffusion over diffusion for extremely long video generation. *arXiv preprint arXiv:2303.12346*, 2023. 3

[48] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. Celebv-text: A large-scale facial text-video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14805–14814, 2023. 6

[49] Rui Zhang, Yixiao Fang, Zhengnan Lu, Pei Cheng, Zebiao Huang, and Bin Fu. Lingualinker: Audio-driven portraits animation with implicit facial control enhancement. *arXiv preprint arXiv:2407.18595*, 2024. 2

[50] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661, 2023. 2, 3, 4, 6

[51] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 6

[52] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pages 650–667. Springer, 2022. 6