EchoMimic: Lifelike Audio-Driven Portrait Animations through Editable Landmark Conditions

Zhiyuan Chen*, **Jiajiong Cao***, **Zhiquan Chen, Yuming Li†, Chenguang Ma†**

Terminal Technology Department, Alipay, Ant Group, Hangzhou, China
{juzhen.czy, xiaoyao.cjj, yingzhen.czq, luoque.lym, chenguang.mcg}@antgroup.com

Abstract

The area of portrait image animation, propelled by audio input, has witnessed notable progress in the generation of lifelike and dynamic portraits. Conventional methods are limited to utilizing either audios or facial key points to drive images into videos, while they can yield satisfactory results, certain issues exist. For instance, methods driven solely by audios can be unstable at times due to the relatively weaker audio signal, while methods driven exclusively by facial key points, although more stable in driving, can result in unnatural outcomes due to the excessive control of key point information. In addressing the previously mentioned challenges, in this paper, we introduce a novel approach which we named EchoMimic. EchoMimic is concurrently trained using both audios and facial landmarks. Through the implementation of a novel training strategy, EchoMimic is capable of generating portrait videos not only by audios and facial landmarks individually, but also by a combination of both audios and selected facial landmarks. EchoMimic has been comprehensively compared with alternative algorithms across various public datasets and our collected dataset, showcasing superior performance in both quantitative and qualitative evaluations. The code and models are available on the project page.

Project Page — https://antgroup.github.io/ai/echomimic

Introduction

The recent advancement in image generation has been greatly advanced by the introduction and effectiveness of Diffusion Models (Rombach et al. 2022; Dhariwal and Nichol 2021; Ho, Jain, and Abbeel 2020). Through rigorous training on large image datasets and a stepwise generation process, these models enable the creation of hyper-realistic images with unprecedented detail. This innovative progress has not only reshaped the field of generative models but has also expanded its application into video synthesis for crafting vivid and engaging visual narratives. In the realm of video synthesis, a significant focus lies in generating human-centric content, notably talking head animations, which involves translating audio inputs into corresponding facial expressions. This task is inherently complex due to the intricate nature and diversity of human facial movements. Conventional methods, despite simplifying the process through constraints such as 3D facial modeling or motion extraction from base videos, often compromise the richness and authenticity of facial expressions.

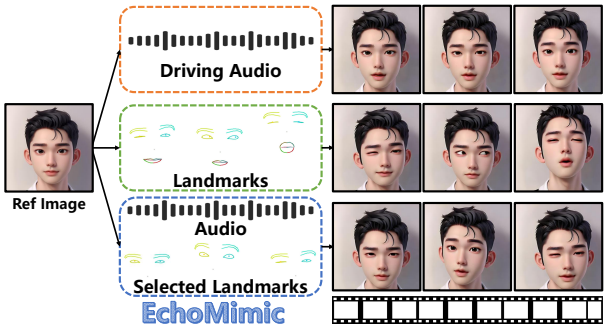

Figure 1: EchoMimic is capable of generating portrait videos by audios, facial landmarks and a combination of both audios and selected facial landmarks.

Portrait animation, a subset of this domain, involves transferring motion and expressions from a source video to a target portrait image using Generative Adversarial Networks (GANs) and diffusion models. Despite the structured two-stage process followed by GAN-based methods (Drobyshev et al. 2022; Liu et al. 2023), which includes feature warping and refinement, they are limited by GAN performance and inaccurate motion depiction, resulting in unrealistic outputs. In contrast, diffusion models have exhibited superior generation capacity, leading to their adaptation for portrait animation tasks. Efforts to enhance these models with specialized modules have been pursued to preserve the portrait's identity and accurately model target expressions. However, challenges such as distortions and artifacts persist, particularly when working with unconventional portrait types, due to inadequate motion representation and inappropriate loss functions for the specific demands of portrait animation. The field faces the dual challenges of synchronizing lip movements, facial expressions, and head poses with audio inputs, and producing visually appealing, high-fidelity animations with consistent temporal coherence. While paramet-

*These authors contributed equally.
†Corresponding authors.

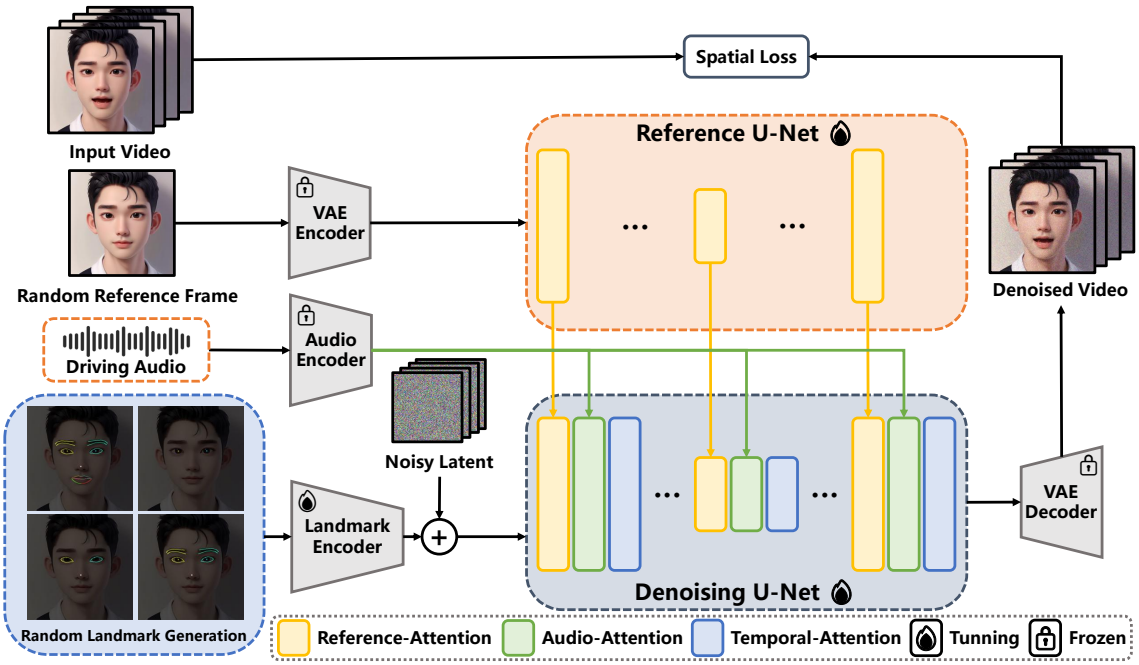

Figure 2: The overall pipeline of the proposed EchoMimic (EM) framework.

ric model-based solutions rely on audio-driven intermediate representations such as 3DMM (Sun et al. 2023), their limitations are imposed by the adequacy of these representations. Decoupled representation learning in the latent space provides an alternative approach by independently addressing the identity and non-identity aspects of facial features. However, it encounters difficulties in achieving comprehensive disentanglement and ensuring consistency across frames.

The quest for progress in portrait animation, particularly talking head image animation, holds substantial significance across various sectors, including gaming, media production, and education. Notable works such as Stable Diffusion (SD) and DiT (Diffusion Models with Transformers) (Peebles and Xie 2023) demonstrate notable advancements in this field. The incorporation of diffusion techniques and parametric or implicit representations of facial dynamics in a latent space facilitates the end-to-end generation of high-quality, realistic animations. However, traditional approaches are constrained to utilizing either audio or facial landmarks for driving images into videos. While these methods can produce satisfactory results, they are associated with specific limitations. For example, approaches driven solely by audio may experience instability due to the relatively weaker audio signal, whereas methods exclusively driven by facial landmarks, although more stable in driving, can lead to unnatural outcomes due to the excessive control of landmark information.

To address the aforementioned challenges, in this paper, we present a novel approach called EchoMimic. EchoMimic is concurrently trained using both audio signals and facial landmarks. Leveraging a novel training strategy, as shown in Figure 1, EchoMimic demonstrates the ability to generate portrait videos using either audios or facial landmarks independently, as well as a combination of audios and selected facial landmarks. EchoMimic is extensively compared with alternative algorithms across diverse public datasets and our collected dataset, demonstrating superior performance in both quantitative and qualitative evaluations.

Related Work

Diffusion Models

Diffusion models have ascended as pivotal generative architectures, showcasing versatility across multimedia tasks (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020). Notably, Score-based Diffusion (SD) models, employing a UNet architecture (Ronneberger, Fischer, and Brox 2015), condition image synthesis on textual descriptions, leveraging large multimodal datasets (Radford et al. 2021). Post-training, these models exhibit flexibility in creative applications, spanning static to dynamic visual media. Innovations, such as Diffusion Models for Text-to-Video Synthesis (DiT) (Peebles and Xie 2023; Esser, Rombach, and Ommer 2021), integrate Transformers with temporal and 3D convolutions, enhancing video generation capabilities. Diffusion models also excel in creating lifelike animated portraits, termed “talking heads”.

Portrait Animation

The evolution of talking head animation has shifted from video-based to image-based approaches, improving realism and expressiveness. Wav2Lip (Prajwal et al. 2020) pioneered audio-driven lip synchronization but faced realism constraints. Recent methods, like Animate Anyone (Hu

2024) and EMO (Tian et al. 2024), leverage diffusion models for enhanced control and consistency in synthesized animations. SadTalker (Zhang et al. 2023) generates 3D motion coefficients from audio for realistic head movement and facial expressions. AniPortrait (Wei, Yang, and Wang 2024) translates audio to detailed 3D facial structures, enriching facial motion. V-Express (Wang et al. 2024) and Hallo (Xu et al. 2024) refine audio-visual synchronization, facial dynamics, and motion variety, achieving high-quality video synthesis.

Despite advances, image-based methods often condition synthesis on audio or pose independently, lacking integrated control. Evaluation metrics favor image quality over facial dynamics, underscoring the need for comprehensive assessments. Addressing these gaps is crucial for advancing the realism and fidelity of talking head animations.

Method

Preliminaries

Our approach is grounded in Stable Diffusion (SD), a seminal framework in text-to-image (T2I) conversion that builds upon the Latent Diffusion Model (LDM) (Rombach et al. 2022). Central to SD is the application of a Variational Autoencoder (VAE) (Kingma and Welling 2013), which acts as an autoencoder. This mechanism transforms the original image's feature distribution, denoted as x_0, into a latent space representation z_0. The encoding phase captures the image essence as $z_0 = E(x_0)$, whereas the decoding counterpart reconstructs it back to $x_0 = D(z_0)$. This design significantly curtails computational expenses without compromising visual quality.

SD integrates principles from the Denoising Diffusion Probabilistic Model (DDPM) (Ho, Jain, and Abbeel 2020) or its variant, the Denoising Diffusion Implicit Model (DDIM) (Song, Meng, and Ermon 2020), introducing a strategic element of Gaussian noise ϵ to the latent representation z_0, yielding a temporally indexed noisy latent state z_t at step t. The inferential phase of SD revolves around a dual objective: progressively eliminating this injected noise ϵ from z_t and concurrently leveraging textual directives. By seamlessly incorporating text embeddings, SD directs the denoising process to yield images that adhere closely to the prescribed textual prompts, thereby realizing finely controlled, high-fidelity visual outputs. The objective function guiding the denoising process during training is formulated as follows:

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

Here, c signifies the text features extracted from the input prompt utilizing the CLIP (Radford et al. 2021) ViT-L/14 text encoder. Within the Stable Diffusion (SD) framework, the estimation of the noise ϵ is accomplished by a customized UNet (Ronneberger, Fischer, and Brox 2015) architecture. This UNet model has been augmented with a cross-attention mechanism, allowing for the effective integration of text features c with the latent representation z_t, thereby enhancing the model's capability to generate images that are coherent with the provided text guidance.

Model Architecture

The core of the proposed EchoMimic framework is the Denoising U-Net architecture, as shown in Figure 2. To improve the network's ability to handle diverse inputs, EchoMimic includes specialized modules: the Reference U-Net for encoding reference images, the Landmark Encoder for using facial landmarks, and the Audio Encoder for processing audio inputs. These components ensure comprehensive encoding, essential for generating high-quality video content.

Denoising U-Net. This architecture enhances multi-frame latent representations corrupted by noise, inspired by the SDv1.5 architecture. It incorporates three attention layers within each Transformer block: the Reference-Attention layer for encoding the relationship between the current frame and reference images; the Audio-Attention layer for capturing the interaction between visual and audio content; and the Temporal-Attention layer for deciphering temporal dynamics between consecutive video frames.

Reference U-Net. The Reference U-Net mirrors the SDv1.5 design and operates in parallel with the Denoising U-Net. Each Transformer block uses a self-attention mechanism to extract reference image features, which are then utilized in the Reference-Attention layer of the Denoising U-Net. The Reference U-Net encodes the reference image without introducing noise, ensuring its essence is accurately captured and seamlessly integrated into the generative process.

Audio Encoder. The Audio Encoder derives the audio representation embedding from the input audio sequence using the pre-trained Whisper-Tiny model (Radford et al. 2023). For each generated frame, the audio features are concatenated with those of adjacent frames to consider temporal context. The Audio-Attention layers in the Denoising U-Net implement a cross-attention mechanism between the latent code and the audio features, ensuring the synthesized character's motion is finely tuned to the accompanying audio.

Landmark Encoder. The Landmark Encoder, implemented as a streamlined convolutional model, encodes each facial landmark image into a feature representation aligned with the latent space dimensions. These features are directly integrated with the multi-frame latents via element-wise addition, enabling the incorporation of precise spatial information critical for maintaining accurate anatomical structure and movement.

Temporal Attention Layer. The Temporal-Attention layers capture the dependencies between successive frames by reshaping the hidden state and applying self-attention mechanisms along the temporal axis. Given a hidden state $h \in R^{b \times f \times d \times h \times w}$, where b, f, d, h, and w denote the batch size, the number of frames, the feature dimension, the height, and the width, respectively, the hidden state is reshaped to $h \in R^{(b \times h \times w) \times f \times d}$. This enables the application of self-attention mechanisms along the temporal axis, ensuring smooth and harmonious transitions in the synthesized frames.

These components collectively enable EchoMimic to produce video sequences with high temporal consistency and

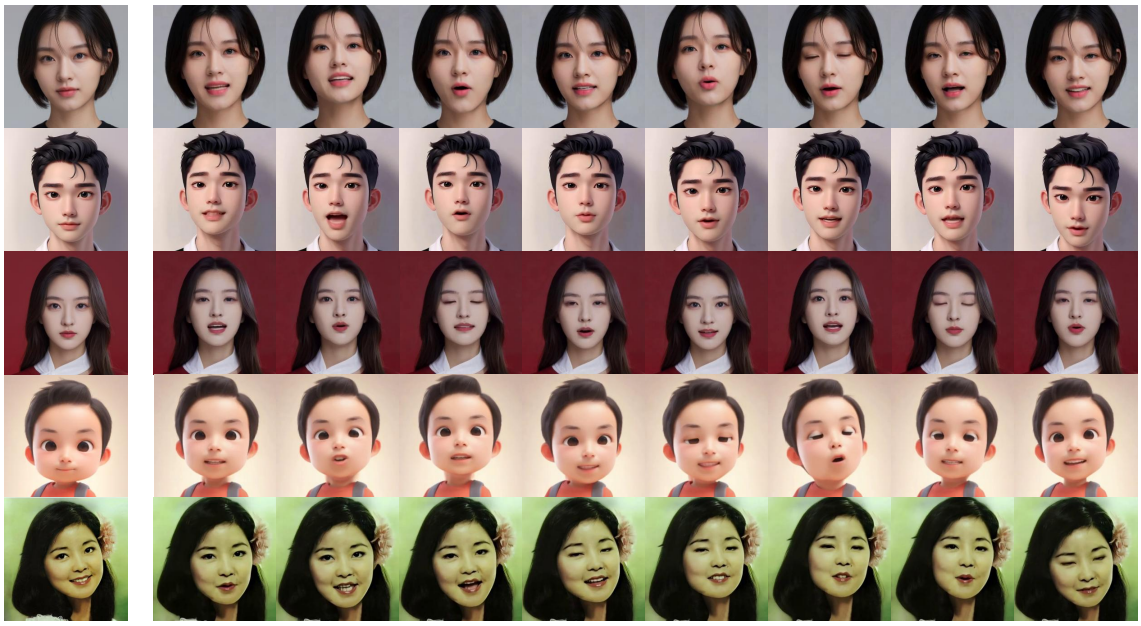

Figure 3: Video generation results of the proposed EchoMimic given different portrait styles and audios.

natural, fluid motion, enhancing the overall visual quality and realism of the generated content.

Spatial Loss. Since the resolution of latent space ($64 * 64$ for $512 * 512$ image) is relative too low to capture the subtle facial details, a timestep-aware spatial loss is proposed to learning the face structure directly in the pixel space. In particular, predicted latent z_t is first mapped to z_0 by sampler. Then the predicted image is obtained via passing z_0 to the vae decoder. Finally, the mse loss is computed on the predicted image and its corresponding ground truth. Besides mse loss, LPIPS loss is adopted to further refine the details of the image. Further, since it is difficult for the model to converge when timestep t is large, we propose a timestep-aware function to reduce the weight for large t. Detailed objective function is shown below:

$$Obj = L_{latent} + \lambda L_{spatial} \tag{2}$$

$$L_{spatial} = w(t)[L2(I_p, I_{GT}) + LPIPS(I_p, I_{GT})] \tag{3}$$

$$w(t) = cosine(t * \pi/2T) \tag{4}$$

Training Details

We adopt a two-stage training strategy following previous works. And we propose efficient techniques including random landmark selection and audio augmentation to boost the training process.

Stage1. In stage1, reference unet and denoising unet are training on single frame data to learning the relations between image-audio and image-pose. In particular, temporal attention layer is not inserted to the denoising unet in stage1.

Method	FID↓	FVD↓	SSIM↑	ED↓	SC↑
SadTalker	41.53	1138.05	0.79	2.24	3.57
AniPortrait	53.14	1038.23	0.75	1.93	4.12
V-Express	58.23	1184.20	0.72	1.80	6.01
Hallo	37.65	501.07	0.78	1.52	**6.64**
EchoMimic	**29.13**	**492.78**	**0.81**	**1.11**	6.37

Table 1: The quantitative comparisons with the existed portrait image animation approaches on the HDTF.

Stage2. In stage2, temporal attention layer is inserted into denoising unet. And the overall pipeline is trained on 12-frame videos for the final video generation. Only the temporal model is trained while the other parts are frozen during stage2.

Random Landmark Selection. To achieve robust landmark-based image driven, we propose a technique called Random Landmark Selection (RLS). In particular, the face is split into several parts including eyebrows, eyes, pupils, nose and mouth. During training, we randomly drops one or several parts of the face.

Spatial Loss and Audio Augmentation. During our experiments, we find that two key techniques can significant improve the quality of the generated video. One is the above proposed spatial loss, which forces the diffusion model to learn the spatial information directly from the pixel space. The other is audio augmentation, which inserts noise and other perturbations to the original audios to achieve similar data augmentations as the images do.

Method	FID↓	FVD↓	SSIM↑	ED↓	SC↑
SadTalker	93.88	1454.32	0.64	3.97	3.71
AniPortrait	92.01	1297.81	0.61	3.91	4.75
V-Express	95.48	2126.24	0.52	4.72	5.79
Hallo	70.42	**1073.71**	**0.64**	2.85	6.07
EchoMimic	**63.25**	1115.85	0.63	**2.72**	**6.14**

Table 2: The quantitative comparisons with the existed portrait image animation approaches on the CelebV-HQ dataset.

Method	FID↓	FVD↓	SSIM↑	ED↓	SC↑
SadTalker	64.63	1681.83	**0.69**	2.15	3.85
AniPortrait	59.69	1544.31	0.66	2.31	4.96
V-Express	62.72	2103.21	0.65	1.68	6.27
Hallo	50.47	1405.21	0.69	1.45	6.75
EchoMimic	**43.27**	**988.14**	0.69	**1.42**	**6.81**

Table 3: The quantitative comparisons with the existed portrait image animation approaches on the our collected dataset.

Inference

For the audio-driven case, the inference process is straightforward. While for the pose-driven or audio+pose-driven case, it is important to align the pose with the reference image according to previous works. Despite there are several techniques proposed for motion alignment, challenges still exist. For instance, existing methods usually apply full face perspective warp affine while ignoring the matching of facial parts. To this end, we propose a developed version of motion alignment called part-aware motion synchronization.

Part-aware Motion Synchronization. Part-aware Motion Synchronization splits the face into several parts. Then, a transformation matrix is first computed on full face. Finally, an extra residual transformation matrix is computed on each part, which will add to the previous matrix to obtain the final matrix.

Experiments

Experimental Setups

Implementation Details. Experiments involved training and inference phases on a high-performance computing setup with 8 NVIDIA A100 GPUs. Training comprised two segments of 30,000 steps each, using a batch size of 4 with 512 × 512 pixel video data. In the second phase, 14 video frames were generated per iteration, incorporating latent variables from the motion module with the initial 2 actual video frames. A constant learning rate of 1e-5 was used. The motion module was initialized with pre-trained Animatediff weights. During training, elements such as the reference image, guiding audio, and motion frames were randomly omitted with a 5% chance. For inference, latent variables perturbed with noise were merged with feature representations from the previous motion frames to ensure sequential coherence.

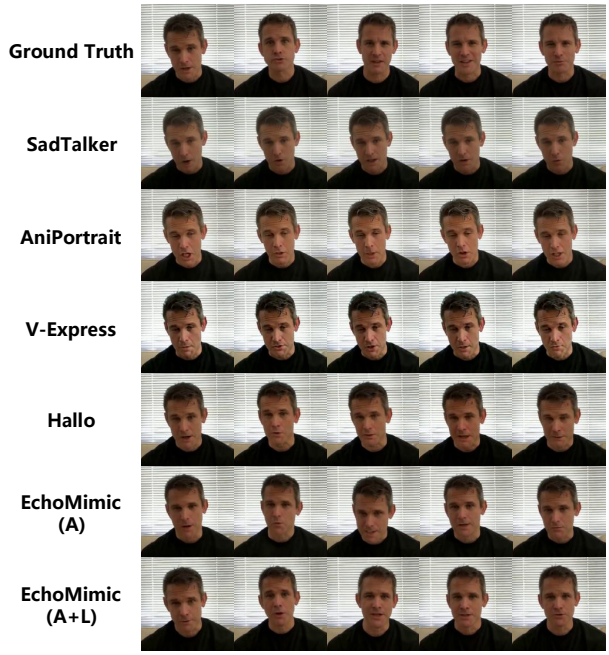

Figure 4: Comparisons with different baseline methods.

Datasets. We collected approximately 540 hours (about 130,000 15-second video clips) of talking head videos, augmented with the HDTF and CelebV-HQ datasets. Data cleaning focused on retaining videos with a single speaker, consistent lip movements, and minimal camera movement. Facial landmarks were extracted using MediaPipe.

Evaluation Metrics. Performance was assessed using FID, FVD, SSIM, E-FID (ED) and Sync-C (SC). FID and FVD measure synthetic image realism, with lower scores indicating better performance. SSIM evaluates structural similarity between ground truth and generated videos. E-FID evaluates facial expression authenticity through face reconstruction and FID calculation of extracted parameters. SyncNet (Chung and Zisserman 2017) is used to calculate Sync-C, which act as indicators of audio-lip synchronization accuracy.

Baseline. Our method was compared against SadTalker (Zhang et al. 2023), AniPortrait (Wei, Yang, and Wang 2024), V-Express (Wang et al. 2024) and Hallo (Xu et al. 2024) across the HDTF, CelebV-HQ, and our collected dataset. A 90:10 split was used for identity data, with 90% for training. Qualitative comparisons evaluated the reference images, audio inputs, and animated outputs.

Quantitative Results

Comparison on HDTF Dataset. Table 1 presents a quantitative evaluation of various portrait animation methods on the HDTF dataset. EchoMimic outperforms others across multiple metrics. Notably, it achieves the lowest FID score (29.13) and FVD score (492.78), indicating superior visual quality and temporal coherence. Additionally, EchoMimic excels in lip synchronization, as evidenced by the highest

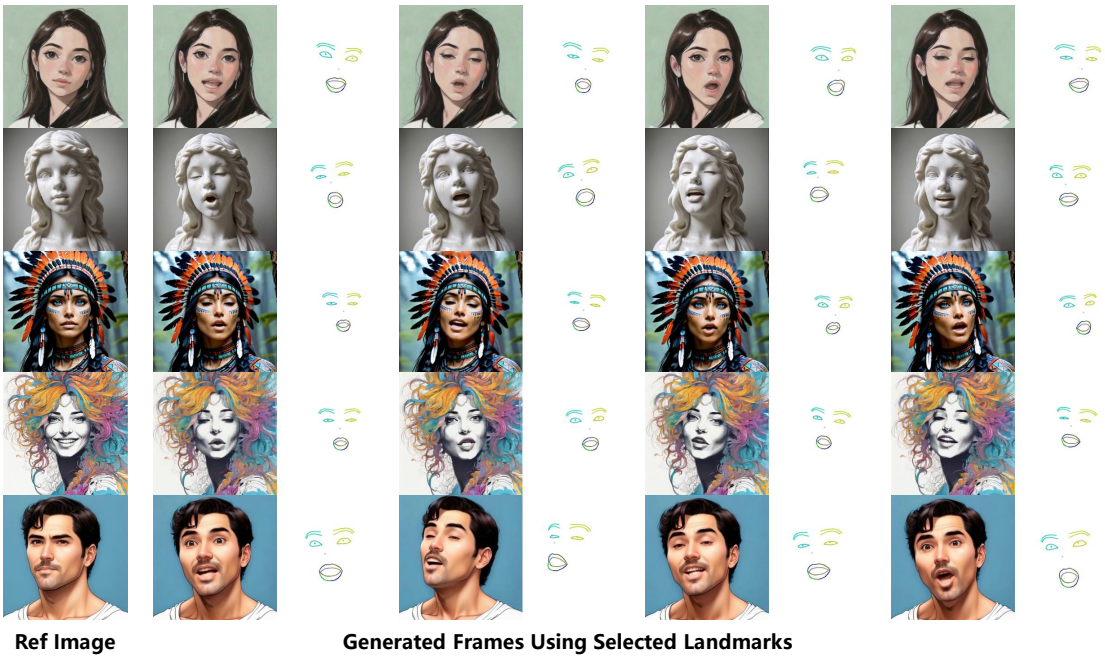

Figure 5: Video generation results of the proposed EchoMimic given different portrait styles and landmarks.

SSIM (0.81) and E-FID (1.11) scores. These results highlight the effectiveness of our method in generating visually compelling and temporally coherent talking head animations with precise lip movement synchronization. For better comparisons, some visual samples are provided in Figure 4.

Comparison on CelebV-HQ Dataset. The CelebV-HQ dataset has lower video quality compared to HDTF and our collected datasets, leading to lower scores across all methods. Table 2 presents a quantitative evaluation using the CelebV-HQ dataset. Our proposed EchoMimic achieves the lowest FID score (63.25) and E-FID score (2.72) among the compared methods, and yields the highest Sync-C score (6.14). These results show its capability to generate high-fidelity animations with excellent temporal consistency.

Comparison on Our Collected Dataset. Table 3 presents the evaluation results for our collected dataset. Our proposed EchoMimic achieves the lowest FID (43.272) and FVD (988.144) scores, indicating superior visual quality and temporal consistency. It also attains a competitive SSIM score (0.691) and the best E-FID score (1.421). Simultaneously, EchoMimic also demonstrates top performance on Sync-C score (6.81) within our collected dataset. These results highlight its capability to generate high-fidelity animations with precise lip synchronization, even under challenging and diverse conditions.

Qualitative Results

Audio Driven. Our audio-driven approach generates high-resolution talking head videos using an audio signal and a reference image. Figure 3 showcases the adaptability and resilience of our method in synthesizing a wide range of audio-visual outputs with seamless synchronization to the accompanying audio. These results affirm its potential for advancing the state-of-the-art in audio-driven video generation.

Landmark Driven. Our landmark-driven approach uses a reference image and landmark controls to generate videos with motion synchronization. Figure 5 displays the qualitative results, demonstrating our method's capability in handling significant pose variations and accurately reproducing nuanced expressions while preserving the identity of the reference portraits.

Audio + Selected Landmark Driven. Audio + selected landmark-driven generation combines an audio signal, a reference image, and selected landmark controls. This mode maintains natural lip synchronization and allows for finer control over facial details, enabling the generated video to exhibit user-desired facial expressions and actions, such as blinking or closing eyes while singing. Figure 6 shows qualitative results with various portrait styles, input audios, and selected landmarks. The results demonstrate clear lip synchronization and precise control over facial expressions consistent with predefined landmarks.

Ablation Study

Facial Landmark Mapping with Motion Synchronization. We validate the efficacy of our motion synchronization method in landmark-driven generation. Figure 7 shows the landmark mapping results with motion synchronization. Compared to previous methods, our proposed motion synchronization aligns landmarks from driving frames with the reference image, facilitating the generation of control results that closely match the face shape in the reference image. For instance, the algorithm projects the small mouth from the reference image onto the significantly larger mouth of the

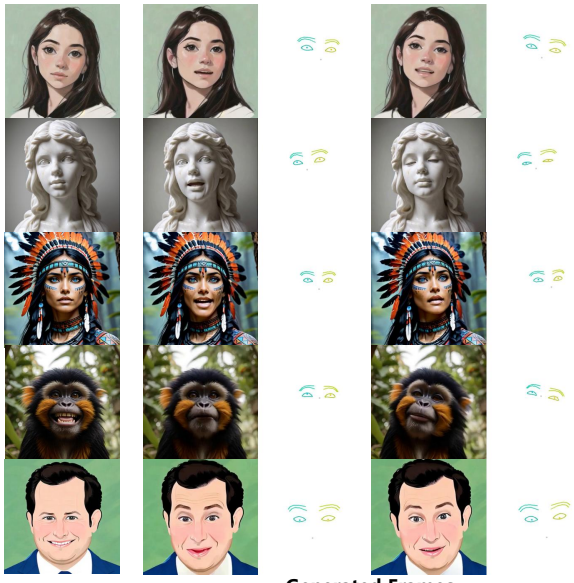

Figure 6: Video generation results of the proposed EchoMimic given different portrait styles, audios and selected landmarks.

Driving Mode	FID↓	FVD↓	SSIM↑	ED↓	SC↑
A	29.13	492.78	0.81	1.11	6.37
L	22.97	156.53	0.88	1.05	6.71
A+L	22.98	181.74	0.88	1.09	6.53

Table 4: Quantitative comparisons of different driving modes of EchoMimic on the HDTF dataset. "A", "L" and "A+L" refer to Audio Only, Pose Only and Audio + Pose, respectively.

potato man, as shown in the figure.

Facial Expression Control by Selected Landmarks. We evaluate the effects of three driving modes: (1) audio driven, (2) facial landmark driven, and (3) audio with selected facial landmark driven. Table 4 provides a quantitative evaluation on the HDTF dataset. The audio-driven mode offers the most freedom, leading to larger disparities from the original videos. The facial landmark-driven mode results in the closest resemblance to the original videos, delivering the best outcomes. The audio with selected facial landmark-driven mode strikes a balance between freedom and similarity, producing intermediate results. Furthermore, we also conduct a comparative analysis between facial landmark driven EchoMimic with other facial landmark driven approaches, such as Follow-Your-Emoji (FYEmoji) (Ma et al. 2024) and LivePortrait (Guo et al. 2024). The experimental results are presented in Table 5. It can be seen that our algorithm also has certain advantages compared to other facial landmark driven algorithms.

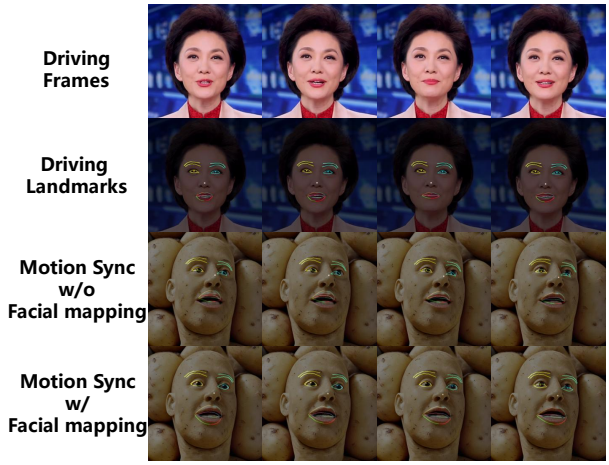

Figure 7: Landmark mapping results with motion synchronization.

Method	FID↓	FVD↓	SSIM↑	ED↓	SC↑
FYEmoji	23.56	168.25	0.85	1.08	6.31
LivePortrait	23.13	**153.27**	**0.89**	1.06	6.58
EchoMimic	**22.97**	156.53	0.88	**1.05**	**6.71**

Table 5: The quantitative comparisons with facial landmark driven approaches on HDTF dataset.

Limitations and Future Work

Update to Video Processing Frameworks. The current architecture is essentially an extension of Stable Diffusion image processing techniques applied to the video domain, rather than a genuine video processing framework. Future work could explore leveraging authentic video processing frameworks (such as 3DVAE, DiT, etc.) to reformulate this method, facilitating enhancements and optimizations tailored specifically for video content (Peebles and Xie 2023; Yu et al. 2023).

Use Acceleration Technique. There is a proliferation of algorithms that accelerate the Stable Diffusion generation process (Chai et al. 2023; Luo et al. 2023). Subsequent research can harness these algorithms to speed up the EchoMimic framework, thereby achieving real-time generation capabilities. This real-time generation can be broadly used for applications such as real-time digital human interactions and conversations.

Conclusions

We presented EchoMimic, a new framework for creating realistic talking head videos using audio and facial landmarks. Our approach outperforms other algorithms in generating authentic and visually appealing animations, as demonstrated through comprehensive evaluations and comparisons. EchoMimic has the potential to enhance multimedia experiences and advance video synthesis. Future work could focus on real-time interactions and expanding its applicability in virtual avatars and entertainment industries.

References

Chai, W.; Zheng, D.; Cao, J.; Chen, Z.; Wang, C.; and Ma, C. 2023. SpeedUpNet: A Plug-and-Play Hyper-Network for Accelerating Text-to-Image Diffusion Models. *arXiv preprint arXiv:2312.08887*.

Chung, J. S.; and Zisserman, A. 2017. 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*, 251–263. Springer.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Drobyshev, N.; Chelishev, J.; Khakhulin, T.; Ivakhnenko, A.; Lempitsky, V.; and Zakharov, E. 2022. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2663–2671.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

Guo, J.; Zhang, D.; Liu, X.; Zhong, Z.; Zhang, Y.; Wan, P.; and Zhang, D. 2024. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, L. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8153–8163.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Liu, H.; Han, X.; Jin, C.; Qian, L.; Wei, H.; Lin, Z.; Wang, F.; Dong, H.; Song, Y.; Xu, J.; et al. 2023. Human motionformer: Transferring human motions with vision transformers. *arXiv preprint arXiv:2302.11306*.

Luo, S.; Tan, Y.; Huang, L.; Li, J.; and Zhao, H. 2023. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*.

Ma, Y.; Liu, H.; Wang, H.; Pan, H.; He, Y.; Yuan, J.; Zeng, A.; Cai, C.; Shum, H.-Y.; Liu, W.; et al. 2024. Follow-Your-Emoji: Fine-Controllable and Expressive Freestyle Portrait Animation. *arXiv preprint arXiv:2406.01900*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. 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*, 484–492.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Sun, X.; Zhang, L.; Zhu, H.; Zhang, P.; Zhang, B.; Ji, X.; Zhou, K.; Gao, D.; Bo, L.; and Cao, X. 2023. Vividtalk: One-shot audio-driven talking head generation based on 3d hybrid prior. *arXiv preprint arXiv:2312.01841*.

Tian, L.; Wang, Q.; Zhang, B.; and Bo, L. 2024. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*.

Wang, C.; Tian, K.; Zhang, J.; Guan, Y.; Luo, F.; Shen, F.; Jiang, Z.; Gu, Q.; Han, X.; and Yang, W. 2024. V-Express: Conditional Dropout for Progressive Training of Portrait Video Generation. *arXiv preprint arXiv:2406.02511*.

Wei, H.; Yang, Z.; and Wang, Z. 2024. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*.

Xu, M.; Li, H.; Su, Q.; Shang, H.; Zhang, L.; Liu, C.; Wang, J.; Van Gool, L.; Yao, Y.; and Zhu, S. 2024. Hallo: Hierarchical Audio-Driven Visual Synthesis for Portrait Image Animation. *arXiv preprint arXiv:2406.08801*.

Yu, L.; Lezama, J.; Gundavarapu, N. B.; Versari, L.; Sohn, K.; Minnen, D.; Cheng, Y.; Gupta, A.; Gu, X.; Hauptmann, A. G.; et al. 2023. Language Model Beats Diffusion–Tokenizer is Key to Visual Generation. *arXiv preprint arXiv:2310.05737*.

Zhang, W.; Cun, X.; Wang, X.; Zhang, Y.; Shen, X.; Guo, Y.; Shan, Y.; and Wang, F. 2023. 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*, 8652–8661.