

HALLO2: LONG-DURATION AND HIGH-RESOLUTION AUDIO-DRIVEN PORTRAIT IMAGE ANIMATION

Anonymous authors

Paper under double-blind review

A APPENDIX

A.1 PRELIMINARIES

Latent Diffusion Models. Latent Diffusion Models (LDMs), introduced by Rombach et al. (2022), represent a significant advancement in generative modeling by conducting diffusion and denoising processes within a compressed latent space rather than directly in the high-dimensional image space. This approach substantially reduces computational complexity while maintaining the quality of generated images.

Specifically, a pre-trained Variational Autoencoder (VAE) Kingma & Welling (2013) is employed to encode input images into lower-dimensional latent representations. Given an input image \mathbf{I} , the encoder $\mathcal{E}(\cdot)$ maps it to a latent vector: $\mathbf{z}_0 = \mathcal{E}(\mathbf{I})$. A forward stochastic diffusion process Sohl-Dickstein et al. (2015); Ho et al. (2020); Song et al. (2020) is then applied to the latent vector \mathbf{z}_0 , adding Gaussian noise over T time steps to produce a sequence of noisy latent variables $\{\mathbf{z}_t\}_{t=1}^T$. The process is defined by: $\mathbf{z}_t = \sqrt{\bar{\alpha}_t} \mathbf{z}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where $t \in \{1, 2, \dots, T\}$ denotes the diffusion steps, $\alpha_t = 1 - \beta_t$ with $\beta_t \in (0, 1)$ being the variance schedule, and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ is the cumulative product of α_t . As t approaches T , the distribution of \mathbf{z}_T converges to a standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ due to the accumulated noise.

The reverse diffusion process aims to reconstruct the original latent vector \mathbf{z}_0 by sequentially denoising \mathbf{z}_T . At each timestep t , a noise prediction network ϵ_θ , typically parameterized using a U-Net architecture Ronneberger et al. (2015), estimates the noise component in \mathbf{z}_t using optional conditioning information \mathbf{c} . The network is trained to minimize the expected mean squared error between the true noise $\boldsymbol{\epsilon}$ and the predicted noise ϵ_θ : $\mathcal{L} = \mathbb{E}_{\mathbf{z}_0, \mathbf{c}, \boldsymbol{\epsilon}, t} [\omega(t) \|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})\|_2^2]$, where $\omega(t)$ is a weighting function that balances the loss contribution across different timesteps.

Once trained, the model can generate new samples by starting from a random Gaussian latent vector $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and iteratively applying the denoising process: $\mathbf{z}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{z}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c}) \right) + \sigma_t \mathbf{n}$, $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, for $t = T, T-1, \dots, 1$, where σ_t is the standard deviation of the noise added at step t . The final latent vector \mathbf{z}_0 is then decoded to reconstruct the image: $\mathbf{I} = \mathcal{D}(\mathbf{z}_0)$, where $\mathcal{D}(\cdot)$ is the decoder of the Variational Autoencoder (VAE).

Incorporating Motion Conditions via Cross-Attention. Incorporating conditioning information is crucial for controlling the generative process in latent diffusion models. Cross-attention mechanisms Vaswani (2017) are employed to effectively integrate motion conditions into the model. The attention layers process both the noisy latent variables \mathbf{z}_t and the embedded motion conditions \mathbf{c} to guide the denoising process. The cross-attention operation is formulated as: $\text{CrossAttn}(\mathbf{z}_t, \mathbf{c}) = \text{softmax}(\mathbf{Q}\mathbf{K}^\top / \sqrt{d_k}) \mathbf{V}$, where $\mathbf{Q} = \mathbf{W}_Q \mathbf{z}_t$, $\mathbf{K} = \mathbf{W}_K \mathbf{c}$ and $\mathbf{V} = \mathbf{W}_V \mathbf{c}$ are the queries; \mathbf{W}_Q , \mathbf{W}_K , and \mathbf{W}_V are learnable projection matrices; and d_k is the dimensionality of the keys. The softmax function ensures that the attention weights sum to one, focusing on the most relevant components of the conditioning information. By integrating cross-attention into the denoising network, the model dynamically adjusts its focus based on the current latent state and the provided conditions. This mechanism enables the generation of images that are coherent with the conditioning inputs, enhancing the expressiveness and realism of the animated portraits.

In our work, the motion conditions \mathbf{c} include the reference image embedding $\mathbf{c}_{\text{image}}$, audio features $\mathbf{c}_{\text{audio}}$, and textual embeddings \mathbf{c}_{text} obtained via Contrastive Language-Image Pretraining

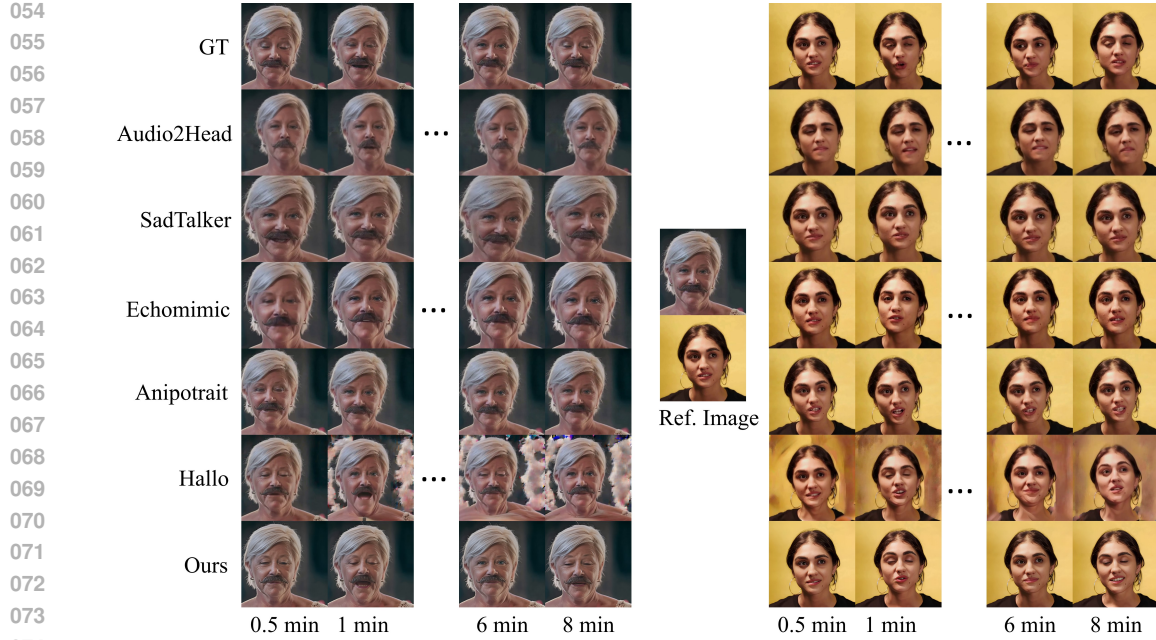


Figure 1: Qualitative comparison with existed approaches on CelebV data-set.

Method	FID↓	FVD↓	Sync-C↑	Sync-D↓	E-FID↓
Audio2Head	57.879	495.421	7.069	7.916	60.538
SadTalker	41.852	588.173	7.026	7.931	21.806
EchoMimic	60.252	805.067	5.499	9.482	19.680
AniPortrait	49.626	583.709	3.810	10.930	22.220
Hallo	82.715	1088.158	6.683	8.420	15.616
Ours	37.944	477.412	6.928	8.307	14.682
Real video	-	-	7.109	7.938	-

Table 1: The quantitative comparisons with existed portrait image animation approaches on the CelebV data-set.

(CLIP) Radford et al. (2021). The combination of these modalities allows for nuanced control over facial expressions, lip movements, and head poses in the generated animations.

A.2 TRAINING AND INFERENCE

Training. This study implements a two-stage training process aimed at optimizing distinct components of the overall framework.

In the initial stage, the model is trained to generate video frames using a reference image, input-driven audio, and a target video frame. During this phase, the parameters of the Variational Autoencoder (VAE) encoder and decoder, as well as those of the facial image encoder, are held constant. The optimization process focuses on the spatial cross-attention modules within both the ReferenceNet and the denoising U-Net, with the objective of enhancing the model’s capabilities for portrait video generation. Specifically, a random image is selected from the input video clip to serve as the reference image, while adjacent frames are designated as target images for training purposes. Additionally, motion modules are introduced to improve the model’s temporal coherence and smoothness.

In the second stage, patch drop and Gaussian noise augmentation techniques are applied to the motion frames to train the model for generating long-duration videos characterized by temporal coherence and smooth transitions. This stage refines the modeling of temporal dynamics by incorporating

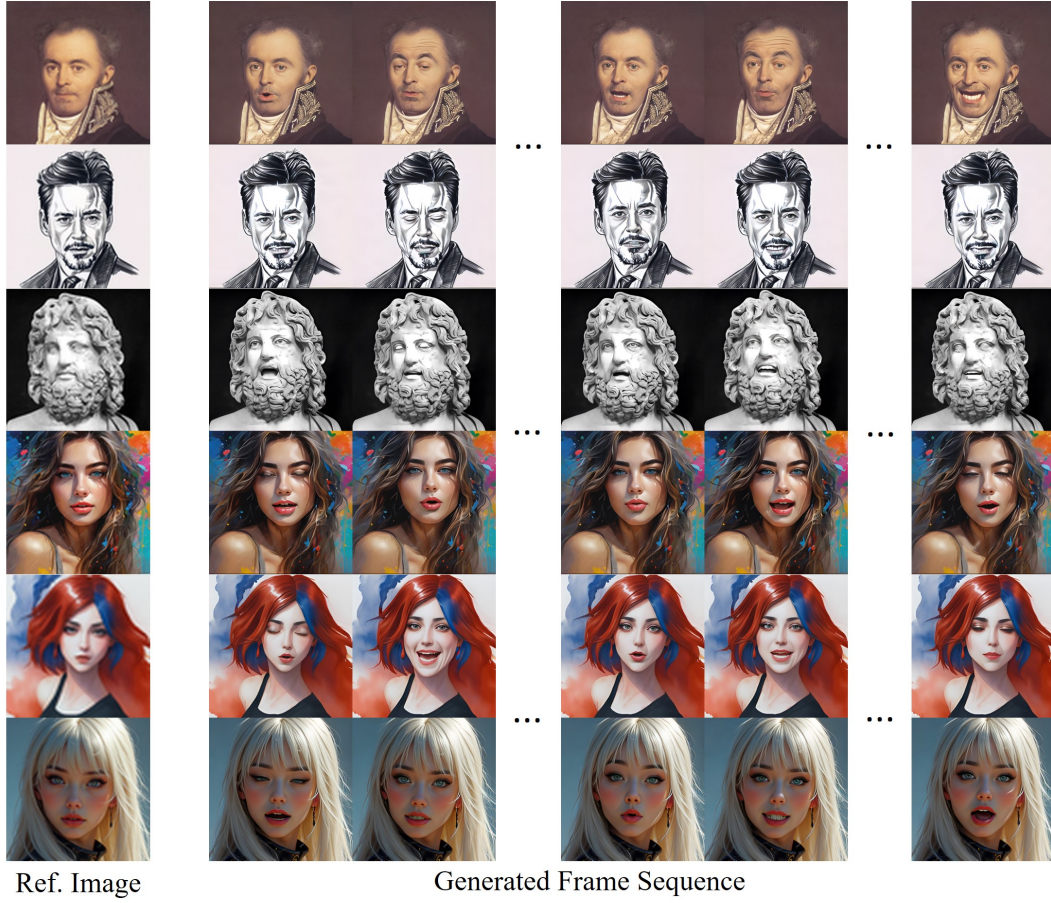


Figure 2: Portrait image animation results given different portrait styles.

corrupted motion frames into the conditioning set, thereby enhancing the model’s ability to capture motion continuity over extended sequences. Concurrently, textual prompts are utilized at this stage to facilitate precise modulation of facial expressions and motions based on textual instructions. For the super-resolution model, the parameters of the VAE encoder are optimized, with a focus on refining the weights responsible for codebook prediction. Temporal alignment is employed within the Transformer-based architecture to ensure consistency and high-quality outputs across frames, thereby enhancing temporal coherence in high-resolution details.

Inference. During inference, the video generation network receives a single reference image, driving audio, an optional textual prompt, and motion frames augmented using patch dropping and Gaussian noise techniques as inputs. The network generates a video sequence that animates the reference image in accordance with the provided audio and textual prompt, synthesizing realistic lip movements and expressions synchronized with the audio output. Subsequently, the high-resolution enhancement module processes the generated video to produce high-resolution frames, thereby enhancing visual quality and fine facial details.

A.3 EXPERIMENTAL SETUPS

Datasets. To evaluate our proposed method, we employed several publicly available datasets, including HDTF, CelebV, and our introduced “Wild” dataset. The “Wild” dataset comprises 2019 clips, totaling approximately 155.9 hours of video content, featuring a diverse array of lip motions, facial expressions, and head poses. This extensive dataset provides a solid foundation for training and testing our portrait image animation framework, facilitating a comprehensive assessment of its ability to generate high-quality and expressive animations across various scenarios.

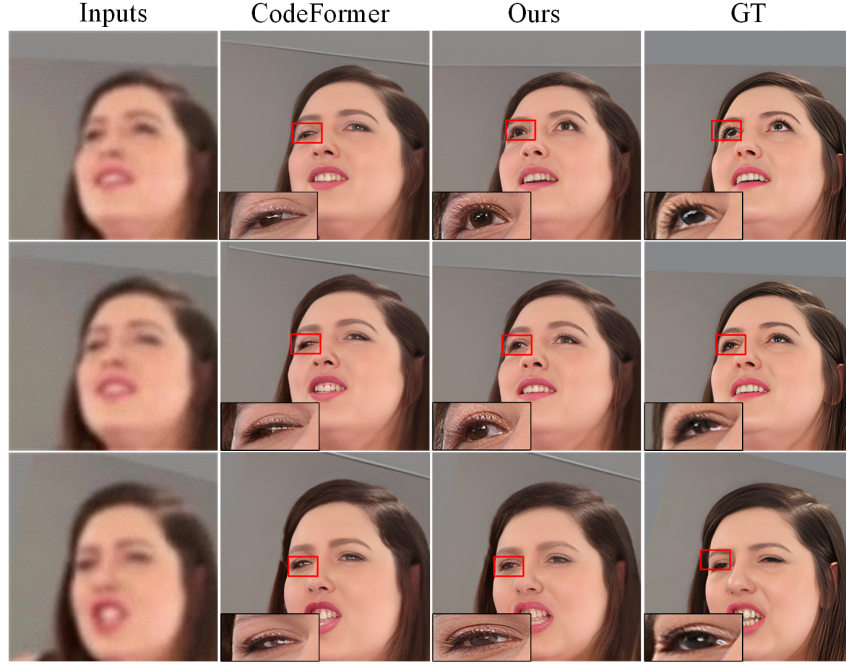


Figure 3: Qualitative comparison between different high-resolution enhancement methods.

Evaluation Metrics. We employ several evaluation metrics to rigorously evaluate our portrait image animation framework. The Fréchet Inception Distance (FID) measures the statistical distance between generated and real images in feature space, with lower values indicating higher quality. The Fréchet Video Distance (FVD) extends this concept to video, assessing the similarity between generated and real videos, where lower values signify superior visual quality. The Sync-C metric gauges lip synchronization consistency with audio, with higher scores reflecting better alignment. Conversely, the Sync-D metric evaluates the temporal consistency of dynamic lip movements, where lower values denote improved motion fidelity. Finally, the Expression-FID (E-FID) quantifies expression synchronization differences between generated content and ground truth videos, providing a quantitative assessment of expression accuracy.

A.4 EXPERIMENTAL RESULTS

Comparison on CelebV Dataset. Table 1 and Figure 1 present the quantitative and qualitative comparisons for the CelebV dataset. Our method achieves the lowest FID of 37.944 and an E-FID of 14.682, indicating superior animation quality. The FVD metric is reported at 477.412, suggesting a coherent video structure. Additionally, our Sync-C score of 6.928 demonstrates competitive performance relative to real video standards. Notably, the increased inference duration has resulted in a significant deterioration in both FID and FVD scores among existing methods, particularly with EchoMimic and Hallo, which exhibit marked degradation in FVD metrics. Additionally, Aniportrait demonstrates notable declines in lip synchronization and expression metrics.

Animation of Different Portrait Styles. Figure 2. This figure illustrates that our method is capable of processing a wide range of input types, including oil paintings, anime images, and portraits from generative models. These findings highlight the versatility and effectiveness of our approach in accommodating different artistic styles.

Comparison between Different High-Resolution Enhancement Methods. Figure 3 provides a qualitative comparison of other image-based enhancement methods. The analysis reveals that integrating super-resolution with temporal alignment significantly enhances visual fidelity, reduces artifacts, and increases image sharpness, resulting in a more coherent and realistic representation of facial features and expressions.

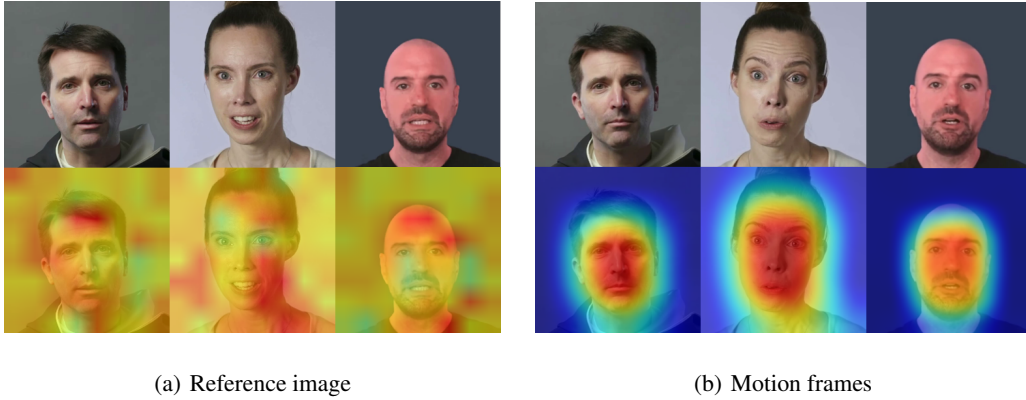


Figure 4: Attention map visualization of the reference image and motion frames.

Attention Map Visualization. Figure 4 presents the attention map visualization, which highlights both the reference image and the temporal attention associated with the motion frames. The results indicate that the reference image indeed influences the overall appearance of the portrait and background due to the implementation of patch drop augmentation. In contrast, the motion frames predominantly focus on regions related to facial motion, underscoring their role in capturing dynamic attributes in the generated animation.

A.5 LIMITATIONS AND FUTURE WORK

Our method for long-duration, high-resolution portrait image animation has several limitations. (1) Reliance on a single reference image constrains the diversity of generated expressions and poses, indicating a need for multiple references or advanced models capable of synthesizing varied facial features. (2) While the patch-drop data augmentation technique effectively preserves motion dynamics, it may introduce artifacts; thus, future research should investigate alternative strategies or adaptive mechanisms for content-specific corruption. (3) The substantial computational demands of generating 4K resolution videos necessitate optimization and hardware acceleration to enable real-time applications.

REFERENCES

- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 2021.
- 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*, 2022.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015.
- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*. PMLR, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.