MegActor-Sigma: Unlocking Flexible Mixed-Modal Control in Portrait Animation with Diffusion Transformer

Shurong Yang*[1], Huadong Li*[1], Juhao Wu*[1], Minhao Jing*[1], Linze Li [1], Renhe Ji †[1], Jiajun Liang †[1], Haoqiang Fan [1], Jin Wang [2]

[1] MEGVII Technology
[2] The University of Hong Kong
{yangshurong6894, jingminhao666}@gmail.com
{lihuadong, wujuhao, lilinze, jirenhe, liangjiajun, fanhaoqiang}@megvii.com

Abstract

Diffusion models have demonstrated superior performance in portrait animation. However, current approaches relied on either visual or audio modality to control character movements, failing to exploit the potential of mixed-modal control. This challenge arises from the difficulty in balancing the weak control strength of audio modality and the strong control strength of visual modality. To address this issue, we introduce MegActor-Sigma: a mixed-modal conditional diffusion transformer (DiT), which can flexibly inject audio and visual modality control signals into portrait animation. Specifically, we make substantial advancements over its predecessor, MegActor, by leveraging the promising model structure of DiT and integrating audio and visual conditions through advanced modules within the DiT framework. To further achieve flexible combinations of mixed-modal control signals, we propose a "Modality Decoupling Control" training strategy to balance the control strength between visual and audio modalities, along with the "Amplitude Adjustment" inference strategy to freely regulate the motion amplitude of each modality. Finally, to facilitate extensive studies in this field, we design several dataset evaluation metrics to filter out public datasets and solely use this filtered dataset for training. Extensive experiments demonstrate the superiority of our approach in generating vivid portrait animations.

Code — https://github.com/megvii-research/megactor.

Introduction

Portrait animation refers to the task of animating a static portrait image using motions and facial expressions from a driving video, while preserving the identity and background of the portrait image. This field has garnered significant attention due to its wide range of applications, including digital avatars (Ma Shugao 2021; Wang, Mallya, and Liu 2021a) and AI-based human conversations (AI 2024; Johnson et al. 2018). In recent years, diffusion models (Xie You 2024; Tian Linrui 2024) have showcased their advantages in portrait animation domain with single-modality control, whether through audio or visual modality.

Figure 1: The visualization of visual leakage. Even when we remove mouth-driven components in visual modality, as V-Express (Wang Cong 2024) does, the generated results still exhibit a certain pattern of speaking without audio-driven.

Previous methods for generating portrait animations with visual modality control (Xu Zhongcong 2023; Zhu Shenhao 2024; Yang Shurong 2024) typically resorted to intermediate motion representations extracted from the driving video, such as landmarks, dense poses, or face meshes. Though demonstrating satisfactory generation quality, obtaining such detailed visual modality control signals in real life scenarios can be inflexible and may require considerable human efforts. Meanwhile, audio-driven methods (Wei, Yang, and Wang 2024; Wang Yuchi 2024; Zhang Yue 2024) allowed users to control portrait animations using audio, supplemented by additional signals like blinking and head rotation. However, these control signals were often insufficient for accurately capturing facial expressions, such as subtle movements of the eyebrows and lips, limiting the model's performance.

To achieve flexible control with superior performance simultaneously, it is essential to endow the model with the ability to freely combine visual and audio modalities as control signals. To this end, the audio modality provides low-cost and flexible lip control, while the visual modality offers accurate motion and facial expression. However, training a model with such comprehensive capabilities poses significant challenges. Previous methods (Wang Cong 2024) have shown that the audio modality often struggles to exert control due to the dominance of the visual modality control. As illustrated in Fig. 1, we observe that even when only partial visual control signals are provided (*i.e.*, eye

*These authors contributed equally.
†Corresponding author

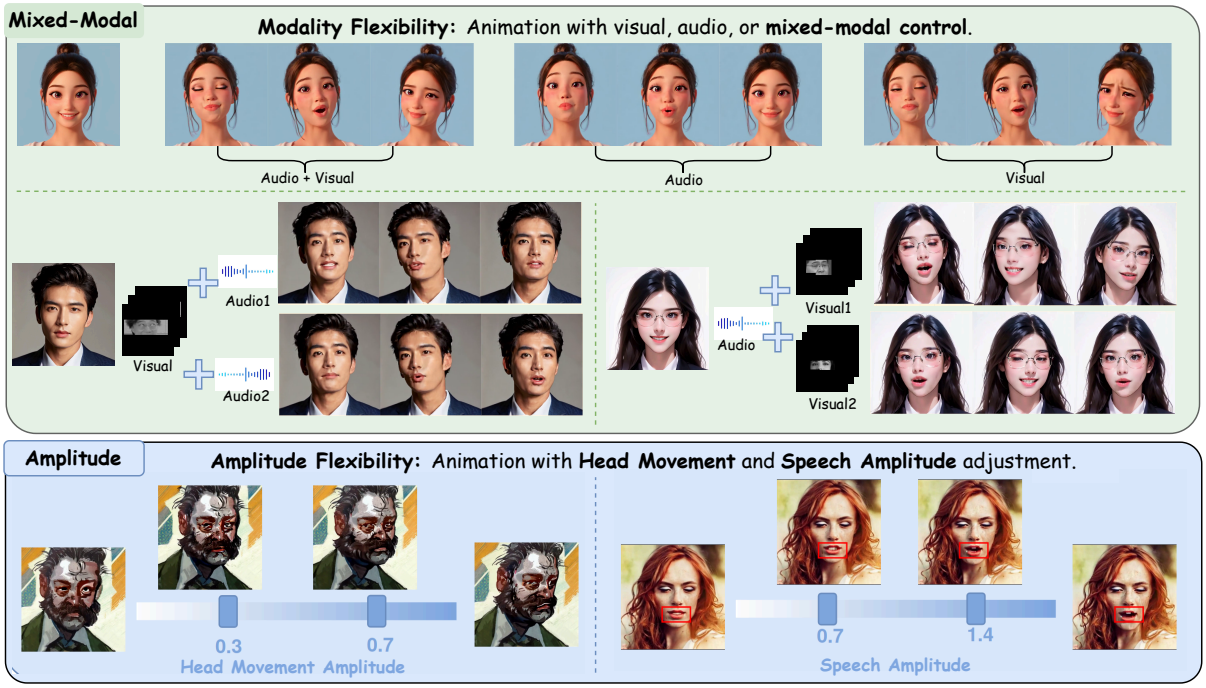

Figure 2: Qualitative results of MegActor-Σ in generating high-quality and flexible portrait animations, include: 1) *Modality Flexibility*, enabling control through visual, audio or mixed-modal control; 2) *Amplitude Flexibility*, enabling adjustment of the scale of head movement and speech amplitude.

patches), the mouth regions still display a speaking pattern (Yang Shurong 2024; Xie You 2024), indicating strong coherence of different facial regions causes visual leakage and prevents the mouth area from following the audio signal cues. Thus, learning a combination of two modalities requires not only the decoupling of modalities, but also the proactive spatial decoupling of visual modality.

To fully explore the potential of decoupled modality control, we present MegActor-Σ, a mixed-modal conditional diffusion transformer (DiT) built upon its predecessor, MegActor (Yang Shurong 2024). Compared to SD1.5, the DiT architecture has fewer parameters, stronger pretrained weights, and produces more stable portrait animations. We further introduce a 3-stage "Modality Decoupling Control" training strategy to prevent visual modality leakage and address the imbalanced control strength of mixed modalities. The first stage, "Spatial-Decoupling Visual Training", aims to spatially decouple visual control. We adopt face dropout strategy to generate spatial patch masks that isolate visual signals, such as those from the eyes or mouth. This method ensures that when eye signals are input, there is no control over the mouth area, thus preventing visual modality leakage. The second stage, "Modality-Decoupling Mixed Training", incorporates audio control while preventing it from being overwhelmed by the visual modality. This involves training with visual, audio and audio-visual mixed signals, while applying face dropout on the visual modality to balance its strength . The final stage, "Motion Priors Training", introduces temporal modules to enable temporal reasoning and consistency. It trains the temporal modules on audio, visual, and mixed modalities without face dropout on the visual modality.

Besides, we introduce a simple yet effective "Amplitude Adjustment" inference strategy to arbitrarily regulate the motion amplitude. For the visual modality, warp transform mapping is used to project the driving images towards a central position, adjusting the magnitude of motion. For the audio modality, the control strength can be adjusted by modifying the combination ratio after the audio attention mechanism, thus enabling customizable speaking amplitude. As shown in Fig. 2, our approach enables flexible modality and amplitude control, resulting in realistic generated videos.

Finally, to facilitate the development of the open-source community in this field, we design several dataset evaluation metrics to filter out public datasets. MegActor-Σ is trained solely on this filtered public dataset, which consists of 313 hours of content featuring high-resolution facial regions, high lip-sync accuracy, and a high proportion of frontal facial orientations. Extensive experiments demonstrate the superior performance of our approach in generating vivid portrait animations, outperforming previous closed-source methods. Our contributions are summarized as follows:

- A novel mixed-modal Diffusion Transformer (DiT) that effectively integrates audio and visual control signals. To the best of our knowledge, this is the first portrait animation method based on the framework of DiT, compared to previous UNet-based methods.
- A novel "Modality Decoupling Control" training strat-

egy that solves visual leakage and effectively balances the control strength between visual and audio modality.

- A set of quality evaluation metrics for filtering public multimodal portrait animation datasets and a filtered 100-hour high-quality dataset for open-source research.
- Extensive experiments demonstrate that our approach excels in generating vivid portrait animations and offers superior flexibility in its application.

Related Work

GAN-based Portrait Animation. Traditional portrait animation methods often employ neural networks to decouple and extract motion features from audio or visual modalities, converting these features into intermediate representations such as landmarks (Su Jiacheng 2024; Song Linsen 2022), 3D head parameters (e.g., 3d Face Morphable Model (Zhang Zicheng 2024; Ma Yifeng 2023; Xing Jinbo 2023; Bai Ziqian 2024; Ran Zimin 2024) (3DMM (Tran and Liu 2018)), Faces Learned with an Articulated Model and Expressions (Ma Haoyu 2024) (FLAME (Li Tianye 2017)), 3D Gaussian parameters (Cho Kyusun 2024; Chen Bo 2024; Li Jiahe 2024; Zhuang Yixiang 2024), 3D Tri-Plane Hash Representation (Li Jiahe 2023)), or latent representations like motion keypoints (Song Luchuan 2024; Gan Yuan 2023; Tan Shuai 2023; Xu Chao 2023; Ji Xinya 2022) and feature encodings (Drobyshev Nikita 2024; Peng Ziqiao 2024; Li Dongze 2024; Su Yaoyu 2024; Liu Yunfei 2023). Techniques such as GANs (Karras, Laine, and Aila 2019) and Gaussian Splatting (Kerbl Bernhard 2023) are then used to render dynamic portrait animations.

Diffusion-based Portrait Animation. Recent advances have seen diffusion models excel in image and video generation (Blattmann Andreas 2023a; Khachatryan Levon 2023; Luo Zhengxiong 2023; Blattmann Andreas 2023b; Guo Yuwei 2023), prompting studies to explore their use in creating high-quality image or videos. Image-driven methods (Hu Li 2023; Chang Di 2023; Wang Tan 2023) use facial landmarks and poses to extract motion for reconstruction but struggle with subtle expressions and can cause facial distortion when the driving signal and reference figure differ in identity and facial proportions, affecting expressiveness and stability. Audio-driven Talking Head generation methods (Xu et al. 2024; Stypułkowski, Vougioukas, and et al. 2024; Xu Sicheng 2024; Tao Liu 2024; Yu Runyi 2024) utilize audio for lip synchronization and weak visual signals for head motion, reducing facial distortion and producing natural videos. However, they fall short in controlling nuanced eye movements and may mismatch head motion amplitude to the driving video, and they often remain uni-modal, missing the benefits of multi-modal control. V-Express (Wang Cong 2024) employs audio and weak image signals for multi-modal control but lacks flexibility due to the simultaneous requirement of both modalities. EchoMimic (Chen Zhiyuan 2024) generates portrait animations under single-modal or multi-modal control. However, it relies on Mediapipe (Lugaresi Camillo 2019) for key points, leading to lost visual detail and high detector dependence. In contrast, MegActor-Σ utilizes the richest original driving images as the control signal and employs an advanced mixed-modal Diffusion Transformer, achieving outstanding performance in portrait animation.

Methodology

The overall framework is summarized in Figure 3. Specifically, MegActor-Σ includes the Denoising Transformer and the Reference Transformer. The Denoising Transformer is designed based on PixArt-α (Chen Junsong 2023), aiming to flexibly infuse the input audio signals and the visual features extracted from the Reference Transformer. The Reference Transformer, which is architecturally identical to the Denoising Transformer, aims to extract fine-grained identity and background details from the reference image. For the driving video, we utilize Driven Encoder, composed of multiple layers of 2D convolutions, to extract motion features. Whisper (Radford Alec 2023) is employed to extract speech audio features, which are then injected into the Denoising Transformer through cross-attention mechanisms. To enhance temporal coherence between generated frames, we integrate a temporal module into the Denoising Transformer and finetune it independently. Based on the framework, we propose a "Modality Decoupling Control" training strategy to balance the dominant visual modality signals with the weak audio modality signals. Besides, to modulate the control motion amplitude across different modalities, we introduce the "Amplitude Adjustment" inference strategy.

MegActor-Σ

Driven Encoder. Inspired by AnimateAnyone (Hu Li 2023), we employ a lightweight Driven Encoder, consisting of four 2D conv layers to extract motion features from the driving video. These motion features are then aligned to the same resolution as the noise latents obtained from random sampling. We concatenate the motion features with the noise latents along the channel dimension. During training, we reinitialize the parameters of the conv-in layer of the Denoising Transformer, retaining the parameters of the first four channels and initializing the remaining channels to zero. This strategy mitigates the disruption to the spatial structure of the Denoising Transformer, originally initialized from PixArt-α, caused by the introduction of motion features.

Spatial Attention Layer. We utilize the Reference Transformer to extract spatial features from the latent representation of the reference image which is yielded by VAE (Kingma and Welling 2013; Rombach et al. 2022). The spatial features are then integrated into the Denoising Transformer via multiple self-attention layers, termed the Spatial Attention, following MagicAnimate (Xu Zhongcong 2023).

Audio Attention Layer. The Denoising Transformer is designed based on PixArt-α, where each block originally contains self-attention layer and cross-attention layer. We replaced the original cross-attention layer with a modified version, termed the audio attention layer, which features a different hidden dimension. The audio attention layer processes audio features encoded by Whisper (Radford Alec 2023).

Temporal Layer. Researches (Guo Yuwei 2023; Xu Jiaqi 2024) have shown that incorporating additional temporal

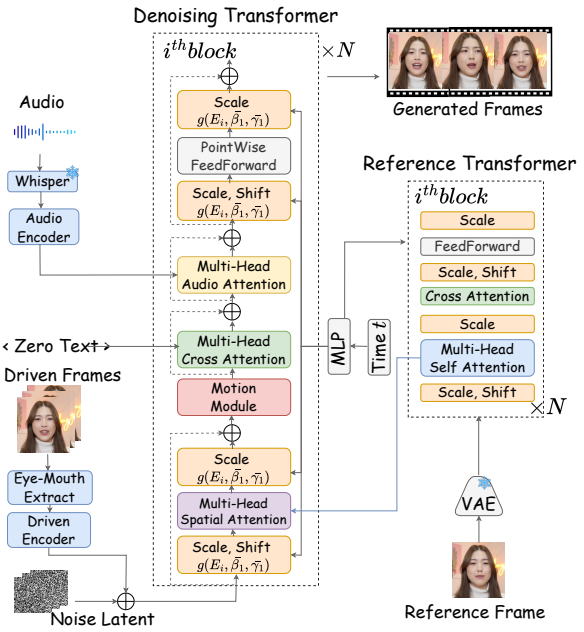

Figure 3: Mixed-modal DiT architecture of MegActor-Σ.

modules into text-to-image (T2I) models for video generation can effectively capture temporal dependencies between frames and enhance their continuity. This design allows for the transfer of pretrained image generation capabilities from the base T2I model. Inspired by these findings (Xu Jiaqi 2024), we integrate a temporal module after each self-attention layer in the Denoising Transformer to facilitate temporal fusion between frames.

Training Strategy

As previously noted in Figure 1, training with mixed-modal control signals often leads to the neglect of audio cues when visual control is applied. We attribute this issue to visual leakage caused by the coherent nature of visual control signals, where the motion of each facial region becomes predictable based on the others. As a result, the audio signals are overshadowed, even when there is no conflicting mouth control signal in the visual modality. Therefore, the key to effective mixed-modal control lies in the spatial decoupling of the eyes control signal and mouth control signal.

To enable flexible combinations of mixed-modal control signals for MegActor-Σ, we propose the "Modality Decoupling Control" training strategy, which can be divided into three stages: "Spatial-Decoupling Visual Training", "Modality-Decoupling Mixed Training" and "Motion Priors Training".

In the "Spatial-Decoupling Visual Training" stage, we use a face dropout strategy to generate spatial masks, randomly selecting isolated control signals, such as those from the eyes, mouth, or both. We then apply these spatial masks to control the regions where attention and the loss function are applied, ensuring that eye signals do not affect the mouth area. In the "Modality-Decoupling Mixed Training" stage, we further integrate audio modality and mixed-modality for combined control. Specifically, the audio modality excludes visual control signals, while the mixed modality comprises audio signals, as in the audio modality, and visual signals that focus solely on the eye region, ensuring that it does not overshadow mouth movements controlled by the audio. This approach aligns the spatial position of the mouth with the audio modality, easing the challenge of modeling mouth shapes. In the "Motion Priors Training" stage, we freeze the attention parameters of the visual and audio modalities and train only the motion module to achieve smooth temporal predictions. We trained our model using the exact same loss functions as PIXART-α, which comprises both KL loss and MSE loss. Detailed information about the training process can be found in Figure 4.

Spatial-Decoupling Visual Training For a video containing facial movements, we randomly select two frames: one as the reference image, I_{ref}, with a static portrait and the other as the predicted result, termed the I_{gt}. We leverage DWPose (Yang Zhendong 2023) to compute facial landmarks for the I_{gt}. During training, random region masks are selected for driving, which can be any combination of the eye and mouth patches M_e, M_m. The driving frame, I_{driven}, will be calculated using the following formula:

$$I_{driven} = M_{driven} \cdot I_{gt}, \text{where } M_{driven} \subseteq \{M_e, M_m\} \tag{1}$$

At this stage of training, the model is initialized from PIXART-α and does not include the audio attention layer, thus no audio information is introduced. The training objective is to reconstruct the I_{gt} based on the I_{driven} and the I_{ref}. During the training process, we introduce an attention mask within the spatial attention mechanism, which is activated only when there is no mouth patch in the driving images. In this case, the embedding of the mouth region will not interact with the driving region through attention, while the rest of the regions remain consistent with the original attention (Chang Di 2023). The mouth region attention formula is as follows:

$$\begin{aligned} Attention(Q, K, V, M_e, M_m) &= \neg M_m \times \sigma(\frac{QK^T}{\sqrt{d}})V \\ &+ M_m \times \sigma(\frac{\neg M_e Q(\neg M_e K)^T}{\sqrt{d}})(\neg M_e V) \end{aligned} \tag{2}$$

where $\sigma(\cdot)$ denotes the Softmax function, and $\neg M$ denotes the complement of the mask M, selecting the inverse regions not covered by the original mask. We also incorporate a spatial mask for MSE loss, which calculates the loss only for the specific regions of interest. For example, when the driving image is eye patch, the attention mask for the mouth tokens during spatial attention excludes the eye area and the MSE loss is computed only for the regions outside of the mouth, as shown in Figure 4. This ensures that the generation of the mouth region is not controlled by the eye region. To further mitigate information leakage arising from content overlap in specific regions between the I_{driven} and I_{gt}, we employ face-swapping (Team 2023) and stylization (Podell Dustin 2023) on the I_{driven}. Furthermore,

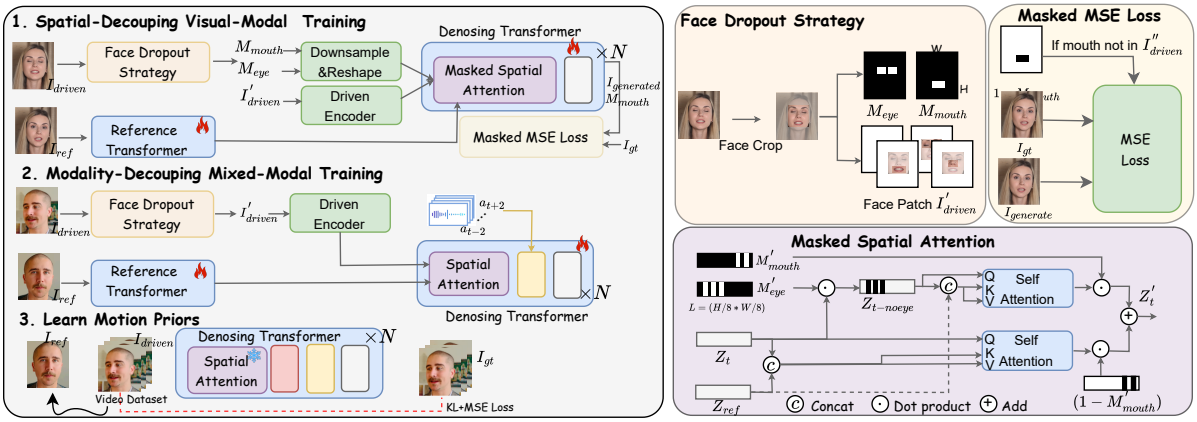

Figure 4: The overall framework of "Modality Decoupling Control" training strategy. Firstly, we utilize face dropout to control partial signals (e.g., eyes or mouth). Masked spatial attention and maksed MSE loss are then applied to ensure that the control of the mouth region is decoupled. Secondly, we integrate audio for mixed-modal control with face dropout to dynamically balance control strength of the audio and visual modalities. Finally, temporal layers are further introduced to learn motion priors.

we perform data augmentation on the driving images, including resolution adjustments, size modifications, and color transformations. During training, we optimize all parameters of the Denoising Transformer, Reference Transformer, and Driven-Encoder.

Modality-Decoupling Mixed Training The model is initialized with the training parameters from the first stage, incorporating an audio attention layer into each transformer block. Using audio signals from the I_{gt} as audio control conditions, the training objective is to reconstruct the I_{gt} based on the I_{driven}, audio conditions. We train all parameters of the Denoising Transformer, Reference Transformer, and Driven-Encoder, while keeping the encoder parameters of Whisper (Radford Alec 2023) frozen. During training, we use 10% of visual modality data with face dropout, 20% of audio modality data and 70% of mixed-modality data.

Motion Priors Training We employ the EasyAnimate-V1(Xu Jiaqi 2024) method to integrate an inter-frame attention module into the model. The training strategy remains the same as in the second stage, with the sole difference being that only the newly inserted motion module is trained.

Inference Strategy

Audio Control For the audio modality, $audio\_scale$ is employed as a weighting factor during the residual summation of outputs from each Multi-head Audio Attention (MHAA), thereby adjusting the influence of audio control.

$$f = f + audio\_scale \times f_{MHAA} \tag{3}$$

As elucidated in Eq. 3, the f is the input and skip feature of the MHAA module, and f_{MHAA} is the output feature of the MHAA module.

Visual control For visual modality control, we first derive the warp transform matrix between the reference image and each frame of the driving video given facial landmarks of all frames. Given the facial landmarks KP_0 of the first frame and KP_i of the i-th frame, M_i is the corresponding warp transform matrix to project KP_0 to KP_i.

$$\begin{aligned} M &= \begin{pmatrix} a, & b, & t_x, \\ c, & d, & t_y \end{pmatrix} \\ \theta_x &= arctan2(b,a), \theta_y = arctan2(d,c) \\ \lambda_x &= \sqrt{a^2+b^2}, \lambda_y = \sqrt{c^2+d^2} \\ M' &= \begin{pmatrix} \lambda_x * cos(\alpha\theta_x), & \lambda_x * sin(\alpha\theta_x), & \alpha t_x, \\ \lambda_y * cos(\alpha\theta_y), & \lambda_y * sin(\alpha\theta_y), & \alpha t_y \end{pmatrix} \end{aligned} \tag{4}$$

As depicted in Eq. 4, we first extract rotation from the transform matrix, and scale rotation together with translation with motion control factor α to obtain a modified Matrix M_i'. Then we warp the i-th frame back using a inverse of M_i and obtained I_i'. We add this extra step to best preserve the look of the original frame. Finally we apply M_i' to I_i' and get the motion-scaled frame for inference. This strategy allows for precise control over the image modality at varying levels of intensity.

Public Data Process

For diffusion models, the quality of the data significantly impacts the effectiveness of the model. However, current state-of-the-art portrait animation methods rely on their private datasets, which hinders community reproducibility and further research. We believe that public data still holds great potential, so we designed a series of evaluation metrics for dataset quality and used them to conduct extensive screening on large-scale multi-modal portrait animation datasets. Ultimately, the models trained on our filtered, high-quality public dataset outperformed those trained on private data. Specifically, we utilized five large-scale multimodal public datasets (Nagrani, Chung, and Zisserman 2017; Chung, Nagrani, and Zisserman 2018; Sung-Bin Kim 2024; Wang, Mallya, and Liu 2021b; Porgali et al. 2023; Hazirbas Caner

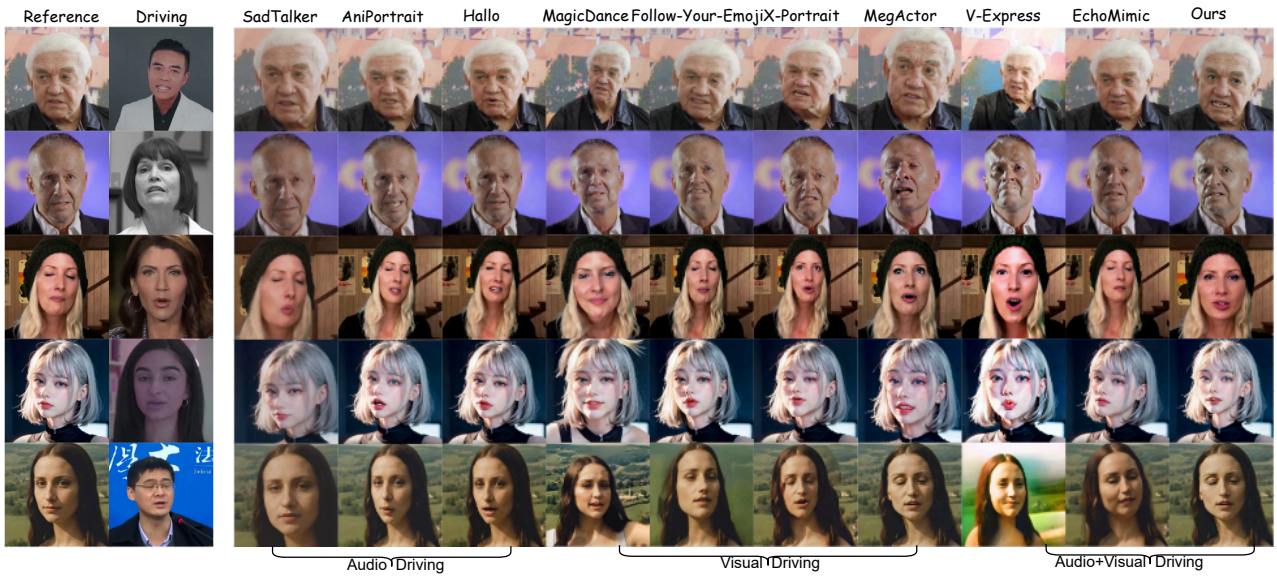

Figure 5: **Qualitative comparisons.** Proposed MegActor-Σ, driven by original images and audio, achieves accurate facial expression transfer (e.g., consistent head and eye movements) and exhibits precise identity resemblance (e.g., facial shape). The reference portraits in rows 1-3 are from VFHQ (Xie Liangbin 2022), and the rest are from the Internet.

Datasets	#F-Hours	Facial Res.↑	Sync-C↑	Sync-D↓	HR Ang.↓
VoxCeleb	140	400 × 400	6.218	8.500	19.4°
Talking-Head	80	400 × 400	4.956	9.315	23.7°
MultiTalk	34	300 × 300	6.002	8.687	20.6°
CCv2	44	600× 600	4.345	9.999	8.2°
HDTF	15	600 × 700	7.957	7.330	5.2°

Table 1: Quantitative comparison of dataset quality. #F-Hours: filtered hours. "Facial Res.": the average resolution of the facial region, "Sync-C" and "Sync-D": lip-audio synchronization accuracy. "HR Ang.": the average head rotation angle of the video clips.

2021), which collectively contained 4,670 hours of raw video footage. After data screening, we retained 313 hours of footage. The evaluation metrics, including effective facial resolution, lip-to-audio synchronization, and head rotate angle, are designed as follows.

Effective Facial Resolution: The quality of the generated dataset is highly dependent on high-resolution facial video dataset. if the face occupies only a small portion of the frame in high-resolution videos, the resolution of the cropped facial images remains low. Therefore, we employed DWPose (Yang Zhendong 2023) to compute the facial bounding box and its size, with a requirement that the resolution of the detected faces must be greater than 600 × 600.

Lip Sync Accuracy: For audio control, precise alignment between audio and lip movement directly impacts the control of mouth region movements by the audio. Therefore, we use SyncNet (Chung and Zisserman 2016) to compute Sync-C and Sync-D, which serve as measures of audio-lip synchronization accuracy. Higher Sync-C values and lower Sync-D values indicate better synchronization. We require that the Sync-C value for the video be greater than 6, and the Sync-D value be less than 8.5.

Head Rotation Angle: The proportion of head rotation angles directly affects the difficulty of training portrait animation methods. While larger head rotation angles can help diversify the dataset, a higher proportion of such data significantly increases the training complexity. Therefore, we calculate the head rotation angle for a video clip based on landmarks, and if the average angle exceeds 30 degrees, the video is excluded.

We calculated the average performance of the datasets across these metrics, which highlights the focus and quality differences among various datasets. Ultimately, we retained 6.7% of the public data for training, as illustrated in Tab. 1.

Experiment Result

Implementation details

The experiments were conducted on 8 NVIDIA V100 GPUs, encompassing both the training and inference phases. Each of the three training stages consisted of 30,000 steps, with the video resolution set to 512 × 512 and a learning rate of 1e-5. Denoising Transformer is composed of 28 basic transformer blocks. To reduce computational complexity and the number of parameters, we inject spatial features from the Reference Transformer into the Denoising Transformer only in the last sixteen blocks. Among the basic transformer blocks numbered 0 through 27, we insert a temporal module only in those with odd-numbered indices.

Method	Framework	Params	Private Data	Modal	HDTF					CCv2				
					FID↓	FVD↓	LPIPS↓	Sync-C↑	Sync-D↓	FID↓	FVD↓	LPIPS↓	Sync-C↑	Sync-D↓
AniPortrait	SD1.5	2.5B	No	A	33.317	576.48	0.1651	4.0886	10.316	37.437	990.78	0.2476	3.6917	10.557
Hallo		2.4B	Yes		35.649	839.22	0.1147	6.7103	8.1199	38.236	988.29	0.1558	6.0820	8.8416
MagicDance	SD1.5	4.1B	No	V	38.399	707.25	0.1861	1.4071	12.833	37.893	794.03	0.2016	1.1953	12.728
MegActor		2.1B	No		32.984	320.42	0.0889	5.8452	8.8823	36.979	402.27	0.1471	5.5078	9.2506
Follow-Your-Emoji		2.2B	Yes		35.063	365.84	0.1090	5.8462	8.5805	36.726	420.90	0.1839	4.8391	9.5500
X-Portrait		2.9B	Yes		32.570	343.52	0.0885	5.4879	9.1072	37.318	491.82	0.1588	5.1656	9.5267
V-Express	SD1.5	2.2B	Yes	A+V	33.770	376.38	0.0960	6.0445	8.4650	36.746	405.98	0.1336	5.6041	9.2320
EchoMimic		2.1B	Yes		31.822	317.04	0.0873	6.2166	8.3247	35.943	388.11	0.1280	6.1575	8.9246
Ours	DiT	**1.4B**	No	A+V	**31.497**	**302.87**	**0.0812**	**6.8226**	**8.1026**	**35.118**	**345.66**	**0.1030**	**6.5407**	**8.6185**

Table 2: Quantitative comparison of our MegActor-Σ with SOTA portrait animation methods on the HDTF and CCv2 Benchmark. Our proposed method achieves superior results with only public dataset training, surpassing those methods trained on private datasets. "A" denotes audio control. "V" denotes visual control. "A+V" denotes the combined of "A" and "V".

Experimental Setup

Evaluation Metrics: The evaluation metrics for portrait animation methods include Fréchet Inception Distance (FID), Fréchet Video Distance (FVD), Learned Perceptual Image Patch Similarity (LPIPS), Synchronization-C (Sync-C) and Synchronization-D (Sync-D). FID, FVD and LPIPS measure the similarity between generated images and real data, with lower values indicating superior performance and more realistic outputs. Sync-C and Sync-D evaluate lip synchronization in generated videos, assessing synchronization from both content and dynamic perspectives.

Baselines: In quantitative experiments, we compared our method with the publicly available implementation of audio-driven methods (Hallo (Xu et al. 2024), AniPortrait (Wei, Yang, and Wang 2024), SadTalker (Zhang and Wenxuan 2023)), image-driven methods (X-Portrait (Xie You 2024), Follow Your Emoji (Ma Yue 2024), MagicDance (Chang Di 2023)), and multi-modal-driven methods (V-Express(Wang Cong 2024), EchoMimic (Chen Zhiyuan 2024)). Evaluations were carried out on HDTF (Zhang Zhimeng 2021) dataset and CCv2 (Porgali et al. 2023) dataset. To ensure a rigorous evaluation, the identity data was partitioned following the standard $9:1$ ratio, with 90% allocated to the training phase.

Evaluations and Comparisons

Table 2 presents comprehensive quantitative evaluations of various portrait animation methods on the HDTF and CCv2 benchmark. Our proposed method demonstrates outstanding performance across multiple metrics, with training on public datasets, outperforming methods trained on private data. Figure 5 illustrates that audio-driven methods struggle to achieve consistency in head movements, eye gaze, etc. Methods driven by landmarks in visual modality and multi-modal approaches (such as MagicDance, Follow-your-emoji, V-Express, and EchoMimic) encounter issues with identity resemblance, as demonstrated in the last column Mona Lisa generation results. Proposed MegActor-Σ combines the accuracy and fidelity of the original images control in the visual modality with the natural smoothness of the audio modality, achieving the most superior results.

Modal	FID↓	FVD↓	LPIPS↓	Sync-C↑	Sync-D↓
A	34.233	383.87	0.1168	6.0315	8.9535
V	32.838	313.33	0.0881	6.4513	8.4818
A+V	**31.497**	**302.87**	**0.0812**	**6.8226**	**8.1026**

Table 3: Quantitative ablation comparison of modalities on HDTF benchamrk. "A" denotes audio control. "V" denotes visual control. "A+V" denotes the combined of "A" and "V".

Ablation Studies

Multi-Modal Control Signals: Table 3 presents the results of uni-modal and multi-modal control tested on the HDTF dataset. The findings indicate that the audio control method (A) imposes the weakest constraints, leading to greater discrepancies from the original video, yet it still achieves high lip synchronization. The method using both audio and image frames (A+V) is constrained by the original image frames and audio, resulting in a generated video that is most similar to the original and attains the highest lip synchronization, thus delivering the best outcomes.

Conclusion

In this paper we have presented MegActor-Σ: a mixed-modal conditional diffusion transformer (DiT) designed to unlock the full potential of versatile mixed-modality control in portrait animation. By addressing the challenges associated with balancing the control strengths of the audio and visual modalities, our method introduces "Modality Decoupling Control" training strategy and "Amplitude Adjustment" inference strategy, enabling more flexible and nuanced control. To further facilitate extensive studies in this field, we design several dataset evaluation metrics to filter out public datasets and solely use this filtered dataset to train MegActor-Σ. Extensive experiments demonstrate the superiority of our approach in generating vivid portrait animations, outperforming previous closed-source methods.

References

AI, D. 2024. https://www.prnewswire.com/news-releases/deepbrain-ai-delivers-ai-avatar-to empower-people-with-disabilities-302026965.html. In *Online*.

Bai Ziqian, e. a. 2024. Efficient 3D Implicit Head Avatar with Mesh-anchored Hash Table Blendshapes. In *CVPR*, 1975–1984.

Blattmann Andreas, e. a. 2023a. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 22563–22575.

Blattmann Andreas, e. a. 2023b. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint*.

Chang Di, e. a. 2023. MagicDance: Realistic Human Dance Video Generation with Motions & Facial Expressions Transfer. *arXiv preprint*.

Chen Bo, e. a. 2024. GSTalker: Real-time Audio-Driven Talking Face Generation via Deformable Gaussian Splatting. *arXiv preprint*.

Chen Junsong, e. a. 2023. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint*.

Chen Zhiyuan, e. a. 2024. EchoMimic: Lifelike Audio-Driven Portrait Animations through Editable Landmark Conditions. *arXiv preprint*.

Cho Kyusun, e. a. 2024. GaussianTalker: Real-Time High-Fidelity Talking Head Synthesis with Audio-Driven 3D Gaussian Splatting. *arXiv preprint*.

Chung, J. S.; Nagrani, A.; and Zisserman, A. 2018. Voxceleb2: Deep speaker recognition. *arXiv preprint*.

Chung, J. S.; and Zisserman, A. 2016. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*.

Drobyshev Nikita, e. a. 2024. EMOPortraits: Emotion-enhanced Multimodal One-shot Head Avatars. In *CVPR*, 8498–8507.

Gan Yuan, e. a. 2023. Efficient emotional adaptation for audio-driven talking-head generation. In *ICCV*, 22634–22645.

Guo Yuwei, e. a. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint*.

Hazirbas Caner, e. a. 2021. Casual conversations: A dataset for measuring fairness in ai. In *CVPR*, 2289–2293.

Hu Li, e. a. 2023. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint*.

Ji Xinya, e. a. 2022. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH*, 1–10.

Johnson, E.; Hervás, R.; Gutiérrez López de la Franca, C.; Mondéjar, T.; Ochoa, S. F.; and Favela, J. 2018. Assessing empathy and managing emotions through interactions with an affective avatar. *Health informatics journal*, 24(2): 182–193.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*, 4401–4410.

Kerbl Bernhard, e. a. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Khachatryan Levon, e. a. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *CVPR*, 15954–15964.

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

Li Dongze, e. a. 2024. Ae-nerf: Audio enhanced neural radiance field for few shot talking head synthesis. In *AAAI*, volume 38, 3037–3045.

Li Jiahe, e. a. 2023. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In *ICCV*, 7568–7578.

Li Jiahe, e. a. 2024. TalkingGaussian: Structure-Persistent 3D Talking Head Synthesis via Gaussian Splatting. *arXiv preprint*.

Li Tianye, e. a. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.*, 36(6): 194–1.

Liu Yunfei, e. a. 2023. Moda: Mapping-once audio-driven portrait animation with dual attentions. In *ICCV*, 23020–23029.

Lugaresi Camillo, e. a. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint*.

Luo Zhengxiong, e. a. 2023. Videofusion: Decomposed diffusion models for high-quality video generation. In *CVPR*, 10209–10218.

Ma Haoyu, e. a. 2024. CVTHead: One-shot Controllable Head Avatar with Vertex-feature Transformer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6131–6141.

Ma Shugao, e. a. 2021. Pixel codec avatars. In *CVPR*, 64–73.

Ma Yifeng, e. a. 2023. Styletalk: One-shot talking head generation with controllable speaking styles. In *AAAI*, volume 37, 1896–1904.

Ma Yue, e. a. 2024. Follow-Your-Emoji: Fine-Controllable and Expressive Freestyle Portrait Animation. *arXiv preprint*.

Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint*.

Peng Ziqiao, e. a. 2024. Synctalk: The devil is in the synchronization for talking head synthesis. In *CVPR*, 666–676.

Podell Dustin, e. a. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint*.

Porgali, B.; Albiero, V.; Ryda, J.; and et al. 2023. The casual conversations v2 dataset. In *CVPR*, 10–17.

Radford Alec, e. a. 2023. Robust speech recognition via large-scale weak supervision. In *ICML*, 28492–28518. PMLR.

Ran Zimin, e. a. 2024. High-Fidelity Facial Albedo Estimation via Texture Quantization. *arXiv preprint*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Song Linsen, e. a. 2022. Audio-driven dubbing for user generated contents via style-aware semi-parametric synthesis. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3): 1247–1261.

Song Luchuan, e. a. 2024. Adaptive Super Resolution for One-Shot Talking-Head Generation. In *ICASSP*, 4115–4119. IEEE.

Stypułkowski, M.; Vougioukas, K.; and et al. 2024. Diffused heads: Diffusion models beat gans on talking-face generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5091–5100.

Su Jiacheng, e. a. 2024. Audio-driven High-resolution Seamless Talking Head Video Editing via StyleGAN. *arXiv preprint*.

Su Yaoyu, e. a. 2024. DT-NeRF: Decomposed Triplane-Hash Neural Radiance Fields For High-Fidelity Talking Portrait Synthesis. In *ICASSP*, 3975–3979. IEEE.

Sung-Bin Kim, e. a. 2024. MultiTalk: Enhancing 3D Talking Head Generation Across Languages with Multilingual Video Dataset. *arXiv preprint*.

Tan Shuai, e. a. 2023. Emmn: Emotional motion memory network for audio-driven emotional talking face generation. In *ICCV*, 22146–22156.

Tao Liu, e. a. 2024. AniTalker: Animate Vivid and Diverse Talking Faces through Identity-Decoupled Facial Motion Encoding. arXiv:2405.03121.

Team, T. M. 2023. ModelScope: bring the notion of Model-as-a-Service to life. https://github.com/modelscope/modelscope.

Tian Linrui, e. a. 2024. EMO: Emote Portrait Alive-Generating Expressive Portrait Videos with Audio2Video Diffusion Model under Weak Conditions. *arXiv preprint*.

Tran, L.; and Liu, X. 2018. Nonlinear 3d face morphable model. In *CVPR*, 7346–7355.

Wang, T.-C.; Mallya, A.; and Liu, M.-Y. 2021a. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, 10039–10049.

Wang, T.-C.; Mallya, A.; and Liu, M.-Y. 2021b. One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing. In *CVPR*.

Wang Cong, e. a. 2024. V-Express: Conditional Dropout for Progressive Training of Portrait Video Generation. *arXiv preprint*.

Wang Tan, e. a. 2023. Disco: Disentangled control for realistic human dance generation. *arXiv preprint*.

Wang Yuchi, e. a. 2024. InstructAvatar: Text-Guided Emotion and Motion Control for Avatar Generation. *arXiv preprint*.

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

Xie Liangbin, e. a. 2022. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *CVPR*, 657–666.

Xie You, e. a. 2024. X-Portrait: Expressive Portrait Animation with Hierarchical Motion Attention. *arXiv preprint*.

Xing Jinbo, e. a. 2023. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *CVPR*, 12780–12790.

Xu, M.; Li, H.; Su, Q.; and et al. 2024. Hallo: Hierarchical Audio-Driven Visual Synthesis for Portrait Image Animation. *arXiv preprint*.

Xu Chao, e. a. 2023. High-fidelity generalized emotional talking face generation with multi-modal emotion space learning. In *CVPR*, 6609–6619.

Xu Jiaqi, e. a. 2024. EasyAnimate: A High-Performance Long Video Generation Method based on Transformer Architecture. *arXiv preprint*.

Xu Sicheng, e. a. 2024. Vasa-1: Lifelike audio-driven talking faces generated in real time. *arXiv preprint*.

Xu Zhongcong, e. a. 2023. Magicanimate: Temporally consistent human image animation using diffusion model. *arXiv preprint*.

Yang Shurong, e. a. 2024. MegActor: Harness the Power of Raw Video for Vivid Portrait Animation. *arXiv preprint*.

Yang Zhendong, e. a. 2023. Effective whole-body pose estimation with two-stages distillation. In *ICCV*, 4210–4220.

Yu Runyi, e. a. 2024. Make Your Actor Talk: Generalizable and High-Fidelity Lip Sync with Motion and Appearance Disentanglement. *arXiv preprint*.

Zhang; and Wenxuan, e. a. 2023. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *CVPR*, 8652–8661.

Zhang Yue, e. a. 2024. MuseTalk: Real-Time High Quality Lip Synchorization with Latent Space Inpainting. *arxiv*.

Zhang Zhimeng, e. a. 2021. Flow-Guided One-Shot Talking Face Generation With a High-Resolution Audio-Visual Dataset. In *CVPR*, 3661–3670.

Zhang Zicheng, e. a. 2024. Learning Dynamic Tetrahedra for High-Quality Talking Head Synthesis. In *CVPR*, 5209–5219.

Zhu Shenhao, e. a. 2024. Champ: Controllable and consistent human image animation with 3d parametric guidance. *arXiv preprint*.

Zhuang Yixiang, e. a. 2024. Learn2Talk: 3D Talking Face Learns from 2D Talking Face. *arXiv preprint*.