
Playmate: Flexible Control of Portrait Animation via 3D-Implicit Space Guided Diffusion

Xingpei Ma * Jiaran Cai * † Yuansheng Guan * Shenneng Huang Qiang Zhang Shunsi Zhang

Guangzhou Quwan Network Technology

<https://playmate.github.io/Playmate/>

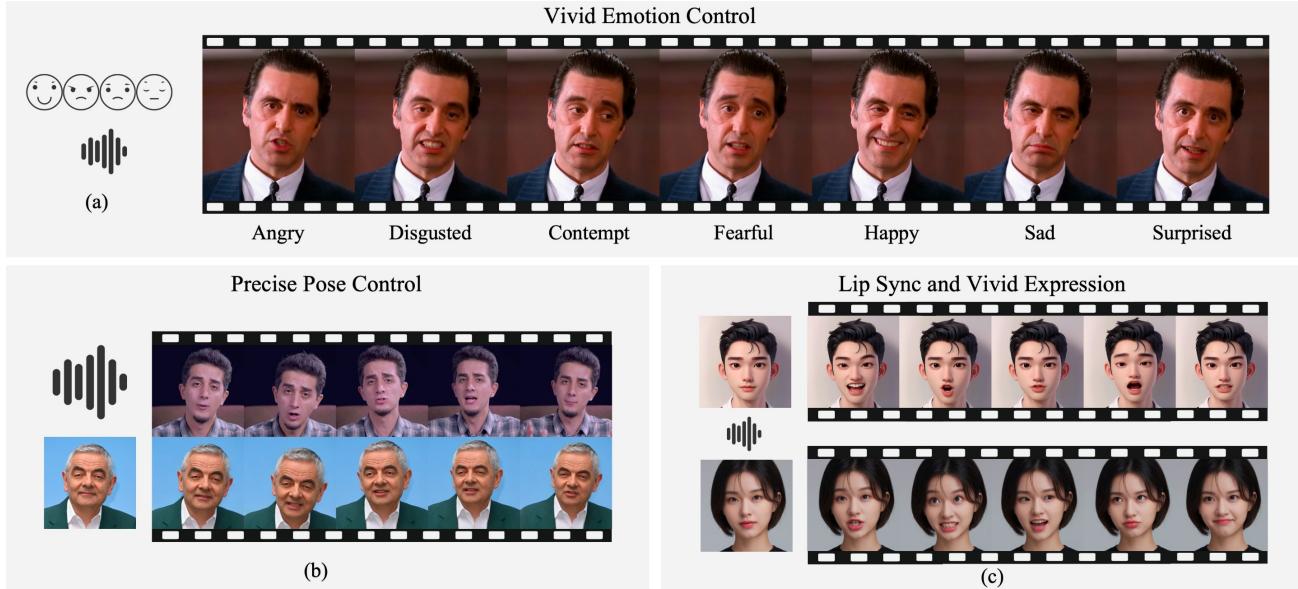


Figure 1. Playmate can generate lifelike talking faces for arbitrary identity, guided by a speech audio clip and a variety of optional control conditions. (a) shows the generation results under different emotional conditions using the same audio clip. The top row in (b) shows the driving images, while the bottom row shows the generated results. The poses in the generated results are controlled by the driving images, and the lip movements are guided by the driving audio. (c) demonstrates highly accurate lip synchronization and vivid, rich expressions across different style images.

Abstract

Recent diffusion-based talking face generation models have demonstrated impressive potential in synthesizing videos that accurately match a speech audio clip with a given reference identity. However, existing approaches still encounter significant challenges due to uncontrollable factors, such as inaccurate lip-sync, inappropriate head

posture and the lack of fine-grained control over facial expressions. In order to introduce more face-guided conditions beyond speech audio clips, a novel two-stage training framework **Playmate** is proposed to generate more lifelike facial expressions and talking faces. In the first stage, we introduce a decoupled implicit 3D representation along with a meticulously designed ***motion-decoupled module*** to facilitate more accurate attribute disentanglement and generate expressive talking videos directly from audio cues. Then, in the second stage, we introduce an ***emotion-control module*** to encode emotion control information into the latent space, enabling fine-grained control over emotions and thereby achieving the ability to generate

*Equal contribution †Project lead . Correspondence to: Jiaran Cai <caijiaran@52tt.com>, Xingpei Ma <maxingpei@52tt.com>.

talking videos with desired emotion. Extensive experiments demonstrate that Playmate not only outperforms existing state-of-the-art methods in terms of video quality, but also exhibits strong competitiveness in lip synchronization while offering improved flexibility in controlling emotion and head pose. The code will be available at <https://github.com/Playmate111/Playmate>.

1. Introduction

Audio-driven portrait animation techniques (Jiang et al., 2024a; Xue et al., 2024) aim to synthesize a lifelike talking face from a static image and speech audio, thereby creating realistic and expressive avatars. In recent years, with the significant development of diffusion-based models (Ho et al., 2020; Song et al., 2020; Dhariwal & Nichol, 2021; Rombach et al., 2022), this field has attracted increasing attention from both academia and industry. Consequently, these advances have unlocked substantial potential for diverse applications, including film dubbing, video generation, and interactive media.

Existing audio-driven portrait animation approaches can be broadly categorized into two main branches: GAN/NeRF-based methods (Chen et al., 2018; Prajwal et al., 2020; Lu et al., 2021; Zhang et al., 2023b; Ma et al., 2023a; Tan et al., 2024b; Ki & Min, 2023; Cheng et al., 2022; Guo et al., 2021b; Wang et al., 2024; Ye et al., 2023) and diffusion-based methods (Shen et al., 2023; Sun et al., 2023; Tian et al., 2025; He et al., 2023; Xu et al., 2024b;a; Zheng et al., 2024; Cao et al., 2024; Ji et al., 2024; Ma et al., 2023b; Chen et al., 2024; Lin et al., 2024b; Liu et al., 2024a; Sun et al., 2024). The former category has primarily focused on the accuracy of lip synchronization (Chung & Zisserman, 2017) and has delivered significant results in this regard. However, GAN/NeRF-based methods often overlook the holistic coordination between audio cues and facial expressions, leading to a failure in generating expressive facial dynamics and lifelike expressions. Recently, the advent of diffusion models has facilitated the generation of high-quality images and videos (Podell et al., 2023; Esser et al., 2024; Liu et al., 2024b; Lin et al., 2024a; Kong et al., 2024; Hong et al., 2022; Yang et al., 2024). Several studies have introduced diffusion models into the field of portrait animation, enabling them to excel in generating talking face videos. Nonetheless, the expression, lip-sync, and head pose of videos generated by diffusion-based methods are strongly correlated and coupled with audio cues, which limits flexibility in controlling specific facial attributes such as head pose and emotional expression, making it challenging to modify one attribute independently without altering the associated audio content. This dependency limits the customization and adaptability of generated animations for different applications.

Based on the above observations, we propose Playmate, a novel two-stage training framework that leverages a 3D-Implicit Space Guided Diffusion Model to generate lifelike talking faces with controllable facial attributes. To achieve this goal, we first introduce a decoupled implicit 3D representation proposed in face-vid2vid (Wang et al., 2021) and LivePortrait (Guo et al., 2024). This implicit representation effectively disentangles multiple face attributes, including expression, lip movement, head pose, and others, thereby enabling more flexible editing and control of these attributes. Subsequently, we train an audio-conditioned diffusion transformer combined with a carefully designed ***motion-decoupled module***, facilitate more accurate attribute disentanglement and generate expressive talking videos directly from audio cues. Finally, to enhance the controllability of emotions, we introduce the ***emotion-control module***, which utilizes DiT blocks (Peebles & Xie, 2023) to encode specified emotion conditions and integrates the encoded information into the aforementioned audio-conditioned diffusion model, thereby achieving flexible control over emotions and improving the editability of the generated talking faces. As shown in Figure 1, Playmate has advantages in audio-driven portrait animation. The main contributions of this paper can be summarized as follows:

- We present Playmate, a novel framework that utilizes 3D-Implicit Space Guided Diffusion for generating talking face videos.
- We meticulously designed a motion-decoupled module and trained a diffusion transformer to improve motion disentanglement and generate motion sequences directly from audio cues.
- A novel emotion-control module and corresponding training strategy are proposed to encode emotion information into latent space and enhance the controllability of emotions of talking head video.
- Experiments show that our method achieves SOTA performance in terms of video quality, motion diversity and emotion controllability.

2. Related Work

2.1. GAN/NeRF-based Audio-driven Portrait Animation

Talking face video generation has been a long-standing challenge in computer vision and graphics. The goal is to synthesize lifelike and synchronized talking videos from driving audio and static reference images. Early GAN/NeRF-based approaches, such as Wav2Lip (Prajwal et al., 2020), Dinet (Zhang et al., 2023b), and VideoReTalking (Cheng et al., 2022), primarily focused on achieving high-quality lip-sync while keeping other facial attributes static. Consequently,

these methods fail to capture strong correlations between the audio and other facial attributes, such as facial expression and head movements. To address this limitation, GANimation (Pumarola et al., 2018) introduced an unsupervised method to generate talking videos with a specific expression. EAMM (Ji et al., 2022) synthesizes emotional talking faces with augmented emotional source videos. More recent studies have typically employed intermediate motion representations (e.g., landmark coordinates, 3D facial mesh, and 3DMM) to generate videos from audio (Zhang et al., 2023a; Gan et al., 2023; Peng et al., 2024). However, such approaches often generate inaccurate intermediate representations, which restricts the expressiveness and realism of the resulting videos. In contrast, our framework generates accurate motion representations based on diffusion transformer.

2.2. Diffusion-based Audio-driven Portrait Animation

Diffusion models have shown impressive performance across various vision tasks. However, previous (Bigioi et al., 2024; Mukhopadhyay et al., 2024; Shen et al., 2023) attempts to utilize diffusion models for generating talking heads have only yielded neutral-emotion expressions, leading to unsatisfactory results. Some of the latest methods have some optimizations for this purpose, such as EMO (Tian et al., 2025), Hallo (Xu et al., 2024a), Echomimic (Chen et al., 2024), and Loopy (Jiang et al., 2024b). EMO introduces a novel framework that ensures consistency in audio-driven animations across video frames, thereby enhancing the stability and naturalness of synthesized speech animations. Hallo contributes a hierarchical audio-driven method for animating portrait images, tackling the complexities of lip synchronization, expression, and pose alignment. MEMO (Zheng et al., 2024) proposes an end-to-end audio-driven portrait animation approach to generate identity-consistent and expressive talking videos. Several of the above methods can generate vivid portrait videos by fine-tuning pre-trained diffusion models. However, they usually use coupled latent spaces to represent facial attributes in relation to the audio. Facial attributes such as expression and head posture are often generated directly from audio cues. This coupling limits the ability to customize control over certain facial attributes, such as pose and expression. In Playmate, we leverage a 3D implicit space that decoupled various facial attributes, enabling diverse and controllable facial animations while maintaining high accuracy in lip synchronization.

2.3. Facial Representation in Audio-driven Portrait Animation

Facial representation learning has been extensively studied in previous works. Various methods (Siarohin et al., 2019; Ren et al., 2021; Li et al., 2017) disentangled variables

using 3DMM, sparse keypoints, or FLAME to explicitly characterize facial attributes. Furthermore, in the field of audio-driven portrait animation, several studies have introduced facial representation techniques to generate lifelike talking videos. Sadtalker (Zhang et al., 2023a) separates generation targets into different categories, including eye blinks, head poses, and lip-only 3DMM coefficients. Recent works such as VASA-1 (Xu et al., 2024b), Takin-ADA (Lin et al., 2024b), DreamTalk (Ma et al., 2023b), and JoyVASA (Cao et al., 2024) have begun to combine face representations with diffusion models to achieve more naturalistic results. Inspired by these advancements, we similarly utilize face representation techniques to generate more natural and controllable talking videos.

3. Methodology

As shown in Figure 2, *Playmate* uses a 3D implicit space as the intermediate representation for generating talking heads from a single static image, guided by a speech audio clip and a set of optional control signals. This section elaborates on our method in detail. We begin with a brief introduction of the 3D implicit space. Furthermore, we describe our meticulously designed approach for generating motion sequences directly from audio cues. Finally, we introduce our emotion-control module and two-stage training strategy, which enhance the controllability of emotions of talking head video.

3.1. Expressive and Disentangled Latent Face Space Construction

Facial representation aims to construct a latent face space that exhibits high degrees of expressiveness and disentanglement. Typically, this approach separates various aspects of facial data into distinct components, such as appearance features and motion attributes. In the field of audio-driven portrait animation, many studies have leveraged these latent spaces to generate talking heads. For example, VASA-1 based its model on the 3D-aid face reenactment framework from (Wang et al., 2021; Drobyshev et al., 2022), while Takin-ADA and JoyVASA constructed their latent space based on face-vid2vid and LivePortrait, respectively. Similarly, we adopt and enhance the decoupled facial representation proposed by face-vid2vid and LivePortrait.

LivePortrait primarily comprises an appearance feature extractor \mathcal{F} , a motion extractor \mathcal{M} , a warping module \mathcal{W} , and a decoder \mathcal{G} . When presented with a source image I_s and a driving image I_d , LivePortrait initially utilizes \mathcal{F} to extract appearance feature f_s from I_s , and separately captures the motion information from both I_s and I_d using \mathcal{M} . Subsequently, it leverages the motion information extracted from I_d to animate I_s using \mathcal{W} and \mathcal{G} , ensuring that the animated result retains the original appearance while adopting the mo-

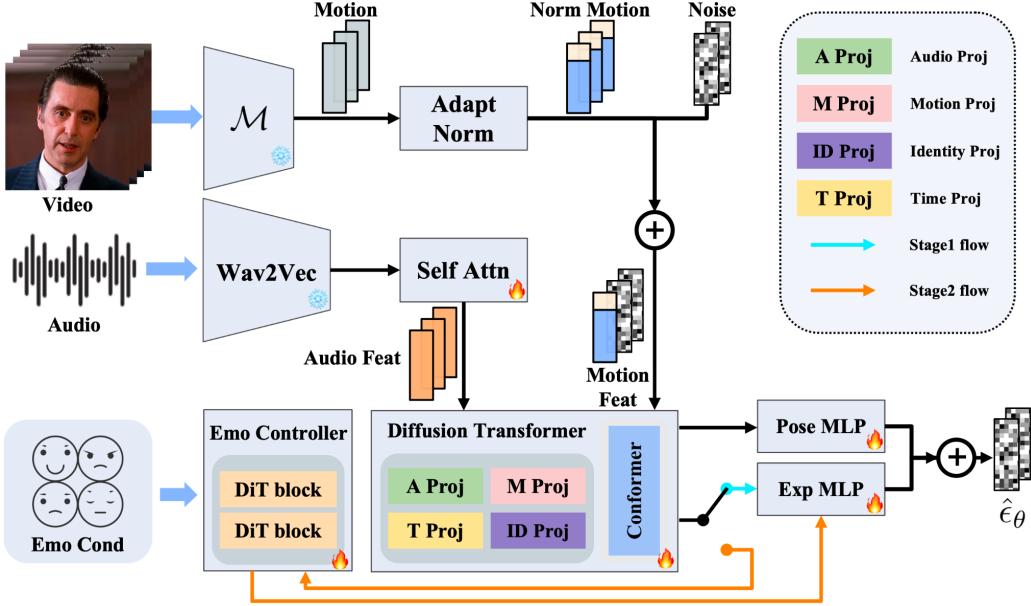


Figure 2. Framework of our approach. Playmate is a two-stage training framework that leverages a 3D-Implicit Space Guided Diffusion Model to generate lifelike talking faces. In the first stage, Playmate utilizes a motion-decoupled module to enhance attribute disentanglement accuracy and trains a diffusion transformer to generate motion sequences directly from audio cues. In the second stage, we use an emotion-control module to encode emotion control information into the latent space, enabling fine-grained control over emotions, thereby improving flexibility in controlling emotion and head pose.

tion cues from the driving image. The motion information is characterized by canonical keypoints $x_c \in \mathbb{R}^{K \times 3}$, expression deformations $\delta \in \mathbb{R}^{K \times 3}$, a rotation matrix $R \in \mathbb{R}^{3 \times 3}$, a translation vector $t \in \mathbb{R}^3$, and a scaling factor $s \in \mathbb{R}$. The transformation in LivePortrait is formalized as:

$$\begin{cases} x_s = s_s \cdot (x_{c,s} R_s + \delta_s) + t_s, \\ x_d = s_d \cdot (x_{c,s} R_d + \delta_d) + t_d, \end{cases} \quad (1)$$

where $x_{c,s}$ represents the canonical keypoints of the source image I_s , and x_s and x_d are the source and driving 3D implicit keypoints after transformation. Additional information about this framework can be found in reference (Guo et al., 2024).

Similar to VASA-1, we found that the original LivePortrait also suffered from poor disentanglement between facial dynamics and head pose. To address this issue, we utilize the pairwise head pose and facial dynamics transfer loss used in VASA-1 to improve its disentanglement. Specifically, let I_i and I_j be two randomly sampled frames from the same video clip. Then we transfer the head pose of I_j onto I_i , resulting in $\hat{I}_{i,j}^{pose} = \mathcal{G}(\mathcal{W}(f_i, x_i, x_{i,j}^{pose}))$, the transformation is formalized as:

$$\begin{cases} x_i = s_i \cdot (x_{c,i} R_i + \delta_i) + t_i, \\ x_{i,j}^{pose} = s_j \cdot (x_{c,i} R_j + \delta_i) + t_j. \end{cases} \quad (2)$$

Similarly, we can transfer I_i 's expression onto I_j , yielding $\hat{I}_{j,i}^{exp} = \mathcal{G}(\mathcal{W}(f_j, x_j, x_{j,i}^{exp}))$, and the transformation is

formalized as:

$$\begin{cases} x_j = s_j \cdot (x_{c,j} R_j + \delta_j) + t_j, \\ x_{j,i}^{exp} = s_j \cdot (x_{c,j} R_i + \delta_i) + t_j. \end{cases} \quad (3)$$

Ultimately, a perceptual loss (Johnson et al., 2016) is employed between $\hat{I}_{i,j}^{pose}$ and $\hat{I}_{j,i}^{exp}$ to enhance the disentanglement between facial dynamics and head pose, the loss function as:

$$\mathcal{L}_p = \| \mathcal{V}(\hat{I}_{i,j}^{pose}) - \mathcal{V}(\hat{I}_{j,i}^{exp}) \|_2, \quad (4)$$

where \mathcal{V} denotes the feature extractor of VGG19 (Simonyan, 2014).

3.2. Audio Conditioned Diffusion Transformer

After constructing the latent face space, we can extract motions by utilizing the frozen \mathcal{M} and train the motion generator, with audio cues serving as the condition.

Adaptive Normalization. To achieve a more effective decoupling of expression and head pose, we employ adaptive normalization by using different means and standard deviations when training the motion generator. Specifically, for expression, we compute the global mean and standard deviation using the entire training dataset, as:

$$\begin{cases} \mu^\delta = \frac{\sum_{i=1}^M \sum_{j=1}^{N_i} \delta_{i,j}}{\sum_{i=1}^M N_i}, \\ \sigma^\delta = \sqrt{\frac{\sum_{i=1}^M \sum_{j=1}^{N_i} (\delta_{i,j} - \mu^\delta)^2}{\sum_{i=1}^M N_i}}, \end{cases} \quad (5)$$

where N_i is the number of frames in the i -th video clip, and M is the total number of all training samples. For head pose, we treat each identity independently and compute the private mean and standard deviation for each sample along the time dimension, as:

$$\begin{cases} \mu_i^\rho = \sum_{j=1}^{N_i} \rho_{i,j}, \\ \sigma_i^\rho = \sqrt{\frac{\sum_{j=1}^{N_i} (\rho_{i,j} - \mu_i^\rho)^2}{N_i}}, \end{cases} \quad (6)$$

where ρ denotes the head pose information. By combining adaptive normalization with the transfer loss mentioned in 3.1, we achieve better decoupling of motion. We refer to this approach as the **motion-decoupled module**.

Speech Representation. Extensive research has demonstrated that pre-trained speech models, such as Wav2Vec2 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021), outperform traditional features like MFCC in performance. Similar to other methods, we utilize Wav2Vec2 as our speech encoder to extract audio features. Additionally, we use a self-attention module to align audio features with motion features.

Diffusion Transformer. The architecture of the Diffusion Transformer is illustrated in Figure 2. We employ four Proj modules, primarily composed of fully connected layers, to extract semantic information from four unique features, aligning this information across different modalities. Next, we utilize a multilayer Conformer (Gulati et al., 2020), a Pose-MLP, and a Exp-MLP in combination with the diffusion formulation to generate motion sequences. Diffusion models define two Markov chains: the forward chain progressively adds Gaussian noise to the target data, while the reverse chain iteratively refines the raw signal from the noise. During training, we gradually transform clean motion m into Gaussian noise m_t following the principles of Denoising Diffusion Probabilistic Models(DDPM) (Ho et al., 2020). The Diffusion Transformer is then trained to reverse this noise-adding process by taking m_t and other conditional features as input and predicting the added noise ϵ . The objective function for training can be expressed as:

$$\mathcal{L}_{diff} = \mathbb{E}_{m_t, f_a, f_{id}, t, \epsilon} \left(\|\epsilon - \hat{\epsilon}_\theta(m_t, f_a, f_{id}, t)\|^2 \right), \quad (7)$$

where f_a , f_{id} are the audio feature and identity feature, $\hat{\epsilon}_\theta$ represents the noise prediction made by the Diffusion Transformer.

3.3. Emotion-control Module

After completing the first training stage, we obtain a diffusion transformer that generates motion sequences guided by audio cues. To enhance the controllability of emotions in talking head videos, we propose the emotion-control module. This involves fixing the parameters of the diffusion

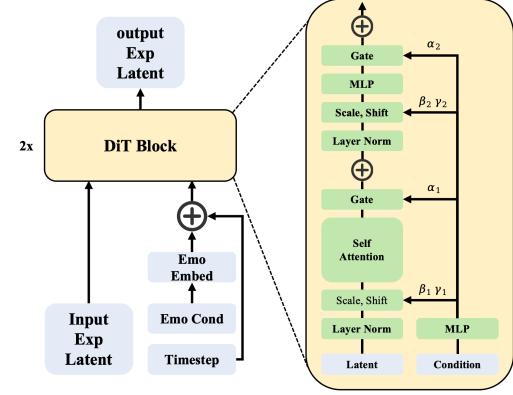


Figure 3. The structure of the Emotion-control Module.

transformer and training an emotion controller based on the DiT block (Peebles & Xie, 2023). As shown in Figure 2 and Figure 3, the emotion-control module consists of two DiT blocks. The first block receives inputs from the conformer and the emotion condition, while the second block takes the output of the first block and emotion condition as its inputs. The outputs of the second block are then passed to Exp-MLP, replacing the original conformer’s output. Through the aforementioned operations, we encode emotion control information into the features and pass them to the Exp-MLP module, thereby achieving the ability to generate talking videos with desired emotion.

During the second phase of training, we freeze the parameters of the diffusion transformer and train only the parameters of the emotion-control module. Experimental results demonstrate that this approach effectively maintains lip-sync accuracy while integrating emotion conditions, allowing for precise emotion control in the generated animations.

Classifier-free guidance(CFG). We adopt a classifier-free guidance (Ho & Salimans, 2022) approach following (Brooks et al., 2023), which has been successfully applied to image generation from multiple conditions. During training, a random dropout strategy is applied to each of the input conditions to improve the model’s robustness and generalization. During inference, we apply:

$$\begin{aligned} \hat{\epsilon}_\theta = & \hat{\epsilon}_\theta(m_t, \emptyset, \emptyset, f_{id}, t) \\ & + w_a [\hat{\epsilon}_\theta(m_t, f_a, \emptyset, f_{id}, t) - \hat{\epsilon}_\theta(m_t, \emptyset, \emptyset, f_{id}, t)] \\ & + w_e [\hat{\epsilon}_\theta(m_t, f_a, f_e, f_{id}, t) - \hat{\epsilon}_\theta(m_t, f_a, \emptyset, f_{id}, t)], \end{aligned} \quad (8)$$

where w_a and w_e are the guidance scales for audio condition and emotion condition, respectively.

Table 1. Quantitative comparisons of video quality and lip synchronization with state-of-the-art methods on two test datasets. The best results are in **bold**, and the second-best are in underlined. Playmate consistently outperforms existing methods in terms of video quality and identity preservation, while also exhibiting strong competitiveness in lip synchronization.

Dataset	Method	FID ↓	FVD ↓	Sync-C ↑	Sync-D ↓	CSIM ↑	LPIPS ↓
HDTF	Hallo (Xu et al., 2024a)	30.484	288.479	7.923	7.531	0.804	0.139
	Hallo2 (Cui et al., 2024)	30.768	<u>288.385</u>	7.754	7.649	0.822	0.138
	MEMO (Zheng et al., 2024)	<u>27.713</u>	299.493	8.059	7.473	<u>0.840</u>	0.132
	Sonic (Ji et al., 2024)	29.189	305.867	9.139	6.549	0.783	0.149
	JoyVASA (Cao et al., 2024)	29.581	306.683	8.522	7.215	0.781	0.157
	Playmate(Ours)	19.138	231.048	8.580	6.985	0.848	0.099
Collected dataset	Hallo (Xu et al., 2024a)	46.114	288.415	6.454	8.384	0.767	0.139
	Hallo2 (Cui et al., 2024)	46.185	295.532	6.509	8.358	0.761	0.144
	MEMO (Zheng et al., 2024)	39.224	260.498	6.569	8.193	<u>0.782</u>	0.130
	Sonic (Ji et al., 2024)	<u>39.069</u>	<u>254.959</u>	7.972	7.124	0.762	0.139
	JoyVASA (Cao et al., 2024)	50.314	304.621	6.858	8.194	0.713	0.163
	Playmate(Ours)	34.716	227.871	<u>7.125</u>	<u>8.007</u>	0.797	0.128

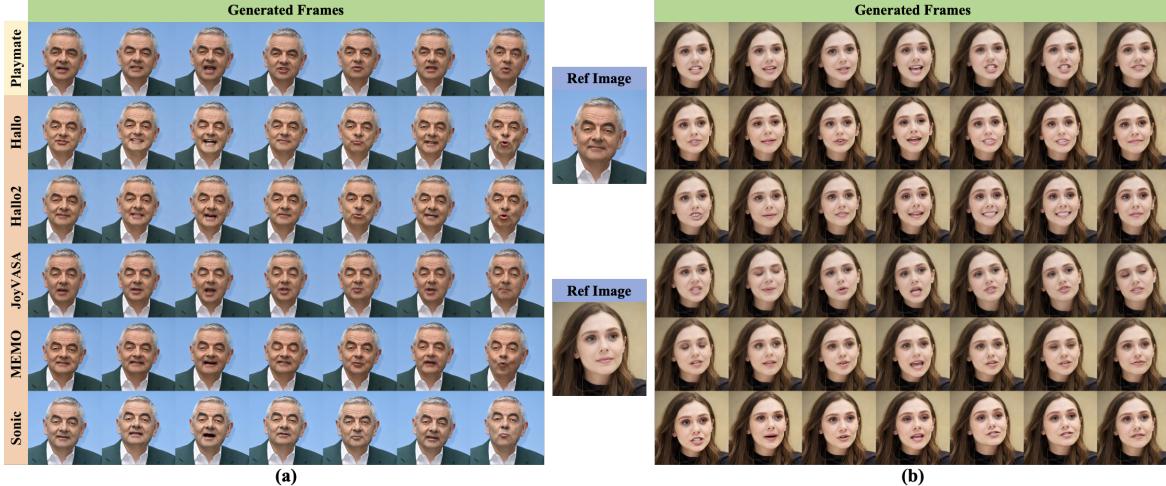


Figure 4. Qualitative comparisons with state-of-the-art methods. Previous methods were prone to generate artifacts in tooth rendering(e.g., (a)-row 6, column 3; (b)-row 3, column 1) and lip synchronization(e.g., (a)-row 4, column 7; (b)-row 2, column 7). Conversely, our approach boasts a superior decoupling capability, which allows it to create more lifelike talking head videos. For more comparison details, please see the Appendix.

4. Experiments

4.1. Experiments Setup

Datasets. We utilize a mixture of datasets, including AVSpeech ([Ephrat et al., 2018](#)), CelebV-Text ([Yu et al., 2023](#)), Acappella ([Montesinos et al., 2021](#)), MEAD ([Wang et al., 2020](#)), MAFW ([Liu et al., 2022](#)), and a talking video dataset collected by us to train our Playmate. In the first stage, we selected approximately 80,000 video clips from the AVSpeech, CelebV-Text, Acappella, and our own dataset to train the diffusion transformer. For the second phase, we selected approximately 30,000 emotionally labeled video clips from the MEAD, MAFW, and our own dataset to train the emotion control module. The duration of each training video ranges from 3 to 30 seconds. We set aside a portion of video clips from our dataset that were not involved in the training process as our out-of-distribution test set. Besides,

we choose HDTF ([Zhang et al., 2021](#)) as our out-of-domain dataset , which comprises about 362 different videos with original resolution of 720P or 1080P.

Implementation Details. In our experiments, the videos are initially converted to 25 fps and subsequently cropped to a resolution of 256×256 pixels based on face landmarks extracted using InsightFace ([Deng et al., 2020; Guo et al., 2021a](#)). The final output resolution is set to 512×512 pixels. During preprocessing, the audios were resampled to 16kHz. The first training phase utilized four NVIDIA A100 GPUs over a 3-day period, with models initialized from scratch. In the second phase, we continued training for two days with two NVIDIA A100 GPUs, while freezing the parameters of the diffusion transformer. For all experiments, we employed the Adam optimizer ([Kingma, 2014](#)). In the inference phase, multi-condition CFG is performed. The CFG scales of the

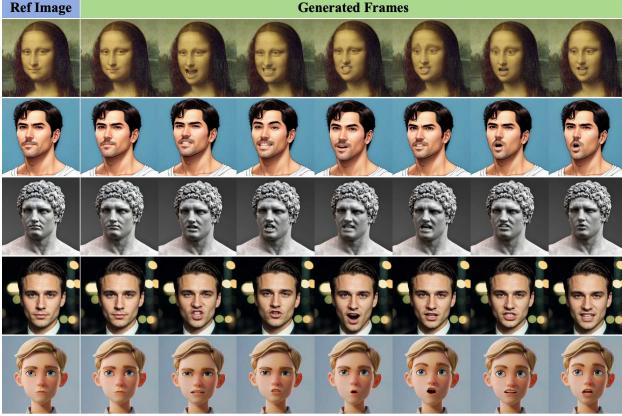


Figure 5. Visualization results in different style images. Playmate can drive a wide range of portraits, including real humans, animations, artistic portraits, and even animals.

audio w_a and the emotion condition w_e are set to 1.5.

Evaluation Metrics. We demonstrate the superiority of our method using multiple widely recognized metrics from previous studies. Specifically, we employ Fréchet Inception Distance (FID) (Heusel et al., 2017) and Fréchet Video Distance (FVD) (Unterthiner et al., 2019) to evaluate the quality of the generated data. For evaluating lip synchronization and motion fluidity, we compute the confidence score (Sync-C) and feature distance (Sync-D) using the pretrained SyncNet (Chung & Zisserman, 2017). Additionally, we compute the cosine similarity (CSIM) of identity vectors extracted using the ArcFace (Deng et al., 2019) face recognition model to evaluate identity preservation. Furthermore, we leverage Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) to measure the feature-level similarity between the generated faces and their ground-truth faces. For the evaluation of emotion control, we employ an additional emotion classifier from (Savchenko, 2023) to calculate the emotion accuracy (Emo-A) of the generated videos.

4.2. Results and Analysis

Quantitative Results. We benchmark our method against SOTA audio-driven portrait animation methods, including Hallo (Xu et al., 2024a), Hallo2 (Cui et al., 2024), JoyVASA (Cao et al., 2024), MEMO (Zheng et al., 2024), and Sonic (Ji et al., 2024). As shown in Table 1, our Playmate significantly outperforms other methods in terms of FID, FVD, CSIM, and LPIPS on two test datasets, while also exhibiting strong competitiveness in lip synchronization. Regarding video quality, our method achieves the lowest FID and FVD scores on both test sets. Specifically, on the HDTF dataset, our FID and FVD scores are 30% and 20% lower than those of the second-best method, respectively, indicating superior video quality compared to other methods. For the CSIM and LPIPS metrics, we also achieve the best results, indi-



Figure 6. Visualization results of emotion control. Each row shows the generation for different identity under different emotional conditions using the same audio clip, demonstrating the flexibility in controlling emotion of Playmate.

cating superior performance in identity preservation and image quality. Additionally, our method achieves good results in Sync-C and Sync-D, both of which are second-best, exhibiting strong competitiveness. Although our quantitative metrics for lip synchronization on the two test sets are marginally lower than Sonic's, they still outperform other compared methods. Furthermore, Sonic is a purely audio-driven algorithm that generates all features of the portrait, including lip movements, expressions, pose, etc., based on audio. This indicates that its driving flexibility is limited. In contrast, we achieve multiple controllable portrait driving methods by constructing a precise attribute disentanglement space, offering users various flexible driving options. The implementation complexity of this decoupling and subsequent driving approach is higher than that of a simple audio-driven method.

Qualitative Results. Figure 4 provides a visualization comparison using open datasets. Upon analysis, previous methods were prone to generate artifacts in tooth rendering(e.g., (a)-row 6, column 3; (b)-row 3, column 1) and lip synchronization(e.g., (a)-row 4, column 7; (b)-row 2, column 7). Conversely, our approach boasts a superior decoupling capability, which allows it to create more lifelike talking head videos. For more comparison details, please see the Appendix. Since images cannot adequately reflect important aspects such as synchronization, naturalness, and stability, the full video comparison will be included in the supplementary materials or on our project page.

Visualization Results in Different Style Images. We further investigate the generation performance of our method

Table 2. Quantitative comparisons of emotion control. The best results are in **bold**, and the second-best are in underlined.

Methods	FID \downarrow	FVD \downarrow	LPIPS \downarrow	Emo-A \uparrow
EAMM (Ji et al., 2022)	111.710	210.275	0.223	0.160
DreamTalk (Ma et al., 2023b)	119.032	199.962	0.246	0.350
EDTalk (Tan et al., 2024a)	135.215	221.897	0.289	0.460
EAT (Gan et al., 2023)	<u>95.085</u>	<u>166.316</u>	<u>0.138</u>	0.450
Playmate(Ours)	68.234	149.837	0.112	0.550

in different style images. Given that our Playmate can directly generate motion sequences from audio, it is capable of driving a wide range of portraits, including real humans, animations, artistic portraits, and even animals. Although this versatility may lead to characters having similar expressions, by integrating the emotion-control module and adjusting the CFG weight, we can easily achieve diverse expressions and movements. As illustrated in Figure 5, our method successfully generates a wide range of expressions and diverse movements across various style images, underscoring its robust performance.

Emotion Control. Figure 6 shows the generation results of our method under different emotional conditions using the same audio clip. These examples clearly illustrate our model’s proficiency in encoding emotional signals into latent space and producing talking face animations with the desired emotion. This highlights Playmate’s effectiveness in achieving precise emotional control while maintaining natural and lifelike results. Furthermore, we compare our method with several existing expression control approaches across various quantitative metrics on the MEAD test dataset. As presented in Table 2, Playmate achieves the highest scores in both video quality and emotion control metrics.

Table 3. User study comparison on open dataset. The best results are in **bold**, and the second-best are in underlined.

Methods	LS \uparrow	VD \uparrow	N \uparrow	VA \uparrow
JoyVASA (Cao et al., 2024)	2.500	2.286	1.714	1.929
Haloo (Xu et al., 2024a)	2.964	2.929	3.071	2.893
Haloo2 (Cui et al., 2024)	3.036	2.929	2.893	2.786
MEMO (Zheng et al., 2024)	3.321	3.036	3.179	3.143
Sonic (Ji et al., 2024)	3.821	<u>3.071</u>	3.750	<u>3.500</u>
Playmate(Ours)	<u>3.750</u>	3.857	<u>3.464</u>	3.643

User Study. We conducted a user study involving 10 participants who rated videos using the MOS (Mean Opinion Score) rating method, on a scale of 1 to 5, across four metrics: Lip Sync (LS), Video Definition (VD), Naturalness (N), and Visual Appeal (VA). As illustrated in Table 3, Playmate has a notable advantage in the VD and VA metrics. While the LS and N metrics are slightly lower than Sonic’s, they still outperform those of other methods, showcasing Playmate’s strong competitiveness.



Figure 7. Ablation study of Adaptive Normalization. Without adaptive normalization, the generated results are prone to pose jittering, inaccurate lip-sync, and strange expressions, resulting in temporal discontinuities and artifacts within the final video.

4.3. Ablation Studies

Adaptive Normalization. We analyze the effectiveness of our proposed Adaptive Normalization in decoupling expressions and head pose. As shown in Figure 7, without adaptive normalization, the generated results are prone to pose jittering, inaccurate lip-sync, and strange expressions, resulting in temporal discontinuities and artifacts within the final video. After introducing adaptive normalization, where head pose feature and expression feature are normalized using different mean and variance values, the model can independently identify and distinguish pose feature and expression feature. Consequently, the final generated videos maintain better continuity in both pose and expression.

CFG scales. By adjusting the CFG scales, we can strike a balance between the quality and diversity of the generated results. In Table 4, we evaluate the selection of CFG scales for the audio and emotion conditions (represented by w_a and w_e in Equation (8)) within our model. Firstly, we conduct ablation experiments solely on w_a , as shown in the upper part of Table 4. When w_a is set to 1.5 or 2.0, Playmate achieves better evaluation metric scores. Upon comprehensive consideration, we set w_a to 1.5 and con-

Table 4. Ablation study of the audio and emotion CFG scales.
The top half shows the ablation results for w_a , and the bottom half shows the ablation results when both w_a and w_e are combined. The best results are in **bold**, and the second-best are in underline.

w_a	w_e	FID \downarrow	FVD \downarrow	Sync-C \uparrow	Sync-D \downarrow	CSIM \uparrow	LPIPS \downarrow	Emo-A \uparrow
1.0	0	17.128	118.481	7.937	7.22	0.939	0.044	0
1.5	0	17.096	<u>120.678</u>	8.141	<u>7.064</u>	0.935	0.046	0
2.0	0	15.968	122.718	8.113	7.049	0.936	0.044	0
2.5	0	<u>16.887</u>	125.738	8.03	7.109	0.936	0.047	0
1.5	1.5	18.948	138.511	7.395	7.644	0.922	0.045	54.405
1.5	2.0	19.978	173.114	<u>7.205</u>	<u>7.931</u>	<u>0.905</u>	0.056	57.579
1.5	2.5	17.498	<u>169.053</u>	7.181	7.933	0.888	<u>0.054</u>	<u>56.276</u>
1.5	3.0	<u>17.825</u>	177.528	7.16	8.075	0.888	0.058	50.055
1.5	3.5	22.32	207.535	6.89	8.313	0.871	0.067	55.753

duct an additional ablation experiments on w_e . To more precisely assess the influence of w_e , we utilize the Emo-A metric, which helps in selecting a suitable value for w_e . As shown in the bottom part of Table 4, when w_a is set to 1.5, setting w_e to 1.5 as well can achieve more balanced results across various testing metrics. Therefore, we also set w_e to 1.5.

5. Conclusion

In summary, we present Playmate, a two-stage training framework designed to generate lifelike talking videos guided by speech audio clips and various optional control conditions. Our approach addresses critical limitations of existing methods, such as precise motion decoupling, expression controllability, and lip synchronization accuracy. In the first stage, Playmate utilizes a motion-decoupled module to enhance attribute disentanglement accuracy and trains a diffusion transformer to produce expressive talking videos directly from audio cues. In the second stage, an emotion-control module is introduced to encode emotion control information into the latent space, enabling fine-grained control over emotions, thereby achieving the ability to generate talking videos with desired emotion. Extensive evaluations show that Playmate consistently surpasses existing state-of-the-art approaches in video quality and facial dynamics realism, while also demonstrating strong performance in lip synchronization and enhanced control over emotion and head pose.

Limitations and future work. While Playmate shows significant advancements, it still has some limitations. Playmate processes information primarily around the face area. Extending its capability to the full upper body or even the whole body could provide additional capability. When using 3D implicit representations, the lack of a more explicit 3D face model may lead to artifacts such as blurred edges during extreme head movements, texture sticking due to neural rendering, and minor inconsistencies in complex backgrounds. Future work will focus on enhancing the model’s robustness to diverse perspectives and styles by incorporating more diverse training data, as well as improving the rendering quality of the framework through advanced techniques.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning in portrait animation in particular. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here. For an extensive discussion of the general ramifications of talking video generation, we point interested readers towards (Xue et al., 2024; Jiang et al., 2024a).

References

- Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.
- Bigioi, D., Basak, S., Stypułkowski, M., Zieba, M., Jordan, H., McDonnell, R., and Corcoran, P. Speech driven video editing via an audio-conditioned diffusion model. *Image and Vision Computing*, 142:104911, 2024.
- Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.
- Cao, X., Shi, S., Zhao, J., Yao, Y., Fei, J., Gao, M., and Wang, G. Joyvasa: Portrait and animal image animation with diffusion-based audio-driven facial dynamics and head motion generation. *arXiv preprint arXiv:2411.09209*, 2024.
- Chen, L., Li, Z., Maddox, R. K., Duan, Z., and Xu, C. Lip movements generation at a glance. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 520–535, 2018.
- Chen, Z., Cao, J., Chen, Z., Li, Y., and Ma, C. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024.
- Cheng, K., Cun, X., Zhang, Y., Xia, M., Yin, F., Zhu, M., Wang, X., Wang, J., and Wang, N. Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–9, 2022.
- Chung, J. S. and Zisserman, A. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20–24, 2016, Revised Selected Papers, Part II 13*, pp. 251–263. Springer, 2017.
- Cui, J., Li, H., Yao, Y., Zhu, H., Shang, H., Cheng, K., Zhou, H., Zhu, S., and Wang, J. Hallo2: Long-duration and high-

- resolution audio-driven portrait image animation. *arXiv preprint arXiv:2410.07718*, 2024.
- Deng, J., Guo, J., Xue, N., and Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.
- Deng, J., Guo, J., Ververas, E., Kotsia, I., and Zafeiriou, S. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5203–5212, 2020.
- Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Neural Information Processing Systems, Neural Information Processing Systems*, Dec 2021.
- Drobyshev, N., Chelishev, J., Khakhulin, T., Ivakhnenko, A., Lempitsky, V., and Zakharov, E. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 2663–2671, 2022.
- Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W. T., and Rubinstein, M. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.
- Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.
- Gan, Y., Yang, Z., Yue, X., Sun, L., and Yang, Y. Efficient emotional adaptation for audio-driven talking-head generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22634–22645, 2023.
- Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.
- Guo, J., Deng, J., Lattas, A., and Zafeiriou, S. Sample and computation redistribution for efficient face detection. *arXiv preprint arXiv:2105.04714*, 2021a.
- Guo, J., Zhang, D., Liu, X., Zhong, Z., Zhang, Y., Wan, P., and Zhang, D. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024.
- Guo, Y., Chen, K., Liang, S., Liu, Y.-J., Bao, H., and Zhang, J. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5784–5794, 2021b.
- He, T., Guo, J., Yu, R., Wang, Y., Zhu, J., An, K., Li, L., Tan, X., Wang, C., Hu, H., et al. Gaia: Zero-shot talking avatar generation. *arXiv preprint arXiv:2311.15230*, 2023.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Hong, W., Ding, M., Zheng, W., Liu, X., and Tang, J. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.
- Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.
- Ji, X., Zhou, H., Wang, K., Wu, Q., Wu, W., Xu, F., and Cao, X. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–10, 2022.
- Ji, X., Hu, X., Xu, Z., Zhu, J., Lin, C., He, Q., Zhang, J., Luo, D., Chen, Y., Lin, Q., et al. Sonic: Shifting focus to global audio perception in portrait animation. *arXiv preprint arXiv:2411.16331*, 2024.
- Jiang, D., Chang, J., You, L., Bian, S., Kosk, R., and Maguire, G. Audio-driven facial animation with deep learning: A survey. *Information*, 15(11):675, 2024a.
- Jiang, J., Liang, C., Yang, J., Lin, G., Zhong, T., and Zheng, Y. Loopy: Taming audio-driven portrait avatar with long-term motion dependency. *arXiv preprint arXiv:2409.02634*, 2024b.
- Johnson, J., Alahi, A., and Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, pp. 694–711. Springer, 2016.
- Ki, T. and Min, D. Stylelipsync: Style-based personalized lip-sync video generation. In *Proceedings of the*

- IEEE/CVF International Conference on Computer Vision*, pp. 22841–22850, 2023.
- Kingma, D. P. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al. Hunyuan-video: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.
- Li, T., Bolkart, T., Black, M. J., Li, H., and Romero, J. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.
- Lin, B., Ge, Y., Cheng, X., Li, Z., Zhu, B., Wang, S., He, X., Ye, Y., Yuan, S., Chen, L., et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024a.
- Lin, B., Yu, Y., Ye, J., Lv, R., Yang, Y., Xie, R., Yu, P., and Zhou, H. Takin-ada: Emotion controllable audio-driven animation with canonical and landmark loss optimization. *arXiv preprint arXiv:2410.14283*, 2024b.
- Liu, T., Chen, F., Fan, S., Du, C., Chen, Q., Chen, X., and Yu, K. Anitalker: animate vivid and diverse talking faces through identity-decoupled facial motion encoding. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6696–6705, 2024a.
- Liu, Y., Dai, W., Feng, C., Wang, W., Yin, G., Zeng, J., and Shan, S. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 24–32, 2022.
- Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., Yuan, Z., Huang, Y., Sun, H., Gao, J., et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024b.
- Lu, Y., Chai, J., and Cao, X. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (ToG)*, 40(6):1–17, 2021.
- Ma, Y., Wang, S., Hu, Z., Fan, C., Lv, T., Ding, Y., Deng, Z., and Yu, X. Styletalk: One-shot talking head generation with controllable speaking styles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1896–1904, 2023a.
- Ma, Y., Zhang, S., Wang, J., Wang, X., Zhang, Y., and Deng, Z. Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv e-prints*, pp. arXiv–2312, 2023b.
- Montesinos, J. F., Kadandale, V. S., and Haro, G. A cappella: Audio-visual singing voice separation. *arXiv preprint arXiv:2104.09946*, 2021.
- Mukhopadhyay, S., Suri, S., Gadde, R. T., and Shrivastava, A. Diff2lip: Audio conditioned diffusion models for lip-synchronization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5292–5302, 2024.
- Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.
- Peng, Z., Hu, W., Shi, Y., Zhu, X., Zhang, X., Zhao, H., He, J., Liu, H., and Fan, Z. SyncTalk: The devil is in the synchronization for talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 666–676, 2024.
- Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- Prajwal, K., Mukhopadhyay, R., Namboodiri, V. P., and Jawahar, C. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 484–492, 2020.
- Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A., and Moreno-Noguer, F. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 818–833, 2018.
- Ren, Y., Li, G., Chen, Y., Li, T. H., and Liu, S. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13759–13768, 2021.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.
- Savchenko, A. Facial expression recognition with adaptive frame rate based on multiple testing correction. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 30119–30129. PMLR, 23–29 Jul

2023. URL <https://proceedings.mlr.press/v202/savchenko23a.html>.
- Shen, S., Zhao, W., Meng, Z., Li, W., Zhu, Z., Zhou, J., and Lu, J. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1982–1991, 2023.
- Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., and Sebe, N. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019.
- Simonyan, K. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- Sun, X., Zhang, L., Zhu, H., Zhang, P., Zhang, B., Ji, X., Zhou, K., Gao, D., Bo, L., and Cao, X. Vividtalk: One-shot audio-driven talking head generation based on 3d hybrid prior. *arXiv preprint arXiv:2312.01841*, 2023.
- Sun, Z., Lv, T., Ye, S., Lin, M., Sheng, J., Wen, Y.-H., Yu, M., and Liu, Y.-j. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics (TOG)*, 43(4): 1–9, 2024.
- Tan, S., Ji, B., Bi, M., and Pan, Y. Edtalk: Efficient disentanglement for emotional talking head synthesis. In *European Conference on Computer Vision*, pp. 398–416. Springer, 2024a.
- Tan, S., Ji, B., and Pan, Y. Flowvqtalker: High-quality emotional talking face generation through normalizing flow and quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26317–26327, 2024b.
- Tian, L., Wang, Q., Zhang, B., and Bo, L. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pp. 244–260. Springer, 2025.
- Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., and Gelly, S. Fvd: A new metric for video generation. 2019.
- Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., and Loy, C. C. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pp. 700–717. Springer, 2020.
- Wang, T.-C., Mallya, A., and Liu, M.-Y. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10039–10049, 2021.
- Wang, X., Ruan, T., Xu, J., Guo, X., Li, J., Yan, F., Zhao, G., and Wang, C. Expression-aware neural radiance fields for high-fidelity talking portrait synthesis. *Image and Vision Computing*, 147:105075, 2024.
- Xu, M., Li, H., Su, Q., Shang, H., Zhang, L., Liu, C., Wang, J., Yao, Y., and Zhu, S. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024a.
- Xu, S., Chen, G., Guo, Y.-X., Yang, J., Li, C., Zang, Z., Zhang, Y., Tong, X., and Guo, B. Vasa-1: Lifelike audio-driven talking faces generated in real time. *arXiv preprint arXiv:2404.10667*, 2024b.
- Xue, H., Luo, X., Hu, Z., Zhang, X., Xiang, X., Dai, Y., Liu, J., Zhang, Z., Li, M., Yang, J., et al. Human motion video generation: A survey. *Authorea Preprints*, 2024.
- Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.
- Ye, Z., He, J., Jiang, Z., Huang, R., Huang, J., Liu, J., Ren, Y., Yin, X., Ma, Z., and Zhao, Z. Geneface++: Generalized and stable real-time audio-driven 3d talking face generation. *arXiv preprint arXiv:2305.00787*, 2023.
- Yu, J., Zhu, H., Jiang, L., Loy, C. C., Cai, W., and Wu, W. Celebv-text: A large-scale facial text-video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14805–14814, 2023.
- Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.
- Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., and Wang, F. 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*, pp. 8652–8661, 2023a.
- Zhang, Z., Li, L., Ding, Y., and Fan, C. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3661–3670, 2021.

Zhang, Z., Hu, Z., Deng, W., Fan, C., Lv, T., and Ding, Y.
Dinet: Deformation inpainting network for realistic face
visually dubbing on high resolution video. In *Proceed-
ings of the AAAI Conference on Artificial Intelligence*,
volume 37, pp. 3543–3551, 2023b.

Zheng, L., Zhang, Y., Guo, H., Pan, J., Tan, Z., Lu, J., Tang,
C., An, B., and Yan, S. Memo: Memory-guided diffusion
for expressive talking video generation. *arXiv preprint
arXiv:2412.04448*, 2024.

A. Appendix

In this appendix, we provide the following materials:

1. More visual comparisons of different methods. Refer to Appendix A.1 in the appendix.
2. More details of Gaussian noise augmentation. Refer to Appendix A.2 in the appendix.
3. Evaluate the accuracy of pose conditional control. Refer to Appendix A.3 in the appendix.
4. Disentanglement of facial representation. Refer to Appendix A.4 in the appendix.



Figure 8. More Visual Comparisons. Each row of images represent the generation results of different methods using the same reference image and the same driving audio.

A.1. More Visual Comparisons

To more comprehensively and objectively demonstrate Playmate’s generation results, we compare the performance of Playmate and several other methods across a wider range of images. As shown in Figure 8, we present a comparison of the effectiveness of Playmate with other current methods in generating talking videos. As evident from the illustration, Hallo and Hallo2 exhibit noticeable artifacts in video quality, characterized by insufficient clarity in teeth generation and inaccurate lip-sync, as exemplified by the images in row 4, column 3 and row 2, column 4 respectively. JoyVASA, despite utilizing a 3D-Implicit Space approach for guidance, falls short in terms of pose and facial motion, resulting in notable head distortions, as exemplified by the images in row 1, column 5 and row 3, column 5. MEMO, on the other hand, struggles with exposure issues, often presenting artifacts on both the face and background, as exemplified by the images in row 2, column 6 and row 7, column 6. Finally, Sonic demonstrates good lip-sync and vivid expressions but still grapples with blurred teeth, as exemplified by the images in row 1, column 7 and row 4, column 7. In sharp contrast, our proposed method showcases the ability to generate more natural facial expressions and head movements that are well-synchronized with audio inputs. Additionally, videos produced by Playmate exhibit superior overall visual quality and stronger identity consistency.

A.2. Gaussian Noise Augmentation

We additionally utilize the speech and motion representation of the previous video frames as input to the audio conditioned diffusion transformer, aiming to maintain the continuity of the generated video. But in this incremental generation process, some contaminations of the previously generated video frames, such as the background noise and subtle distortions in facial expressions, will propagate to subsequent frames and continue to amplify artifacts. In order to enhance the resistance of our diffusion transformer to the above contaminations, we similar to Hallo2 incorporate Gaussian noise into the prev motion representations as:

$$\hat{m}^{prev} = \sqrt{\bar{\alpha}_t} \tilde{m}^{prev} + (1 - \bar{\alpha}_t) \epsilon, \epsilon \sim \mathcal{N}(0, \mathbf{I}), t \in (0, 50), \quad (9)$$

where \tilde{m}^{prev} represents the motion representations of the previous frame, and ϵ represents the Gaussian noise of the corresponding frame. The Gaussian noise follows the normal distribution $\mathcal{N}(0, \mathbf{I})$, in which \mathbf{I} denotes the identity matrix. These noise-augmented motion representations are concatenated with current noisy motions m_t and jointly participate in the diffusion process. Specifically, each denoising step can be described as:

$$\hat{\epsilon}_\theta(m_t, f_a, f_{id}, t) = \hat{\epsilon}_\theta(concat(\hat{m}^{prev}, m_t), f_a, f_{id}, t), \quad (10)$$

where $\hat{\epsilon}_\theta(m_t, f_a, f_{id}, t)$ represents the noise component predicted by Audio Conditioned Diffusion Transformer, and f represents the conditioning inputs of audio feature and identity feature. Through this noise-augmented operation, our diffusion transformer is more robust to slight changes in the motion input, thereby mitigating the impact of contaminations from previously generated video frames.

We conducted an ablation study on Gaussian noise augmentation, as illustrated in Figure 9. Without this augmentation strategy, once artifacts emerged in the generated image, these flaws would propagate, leading to flawed generations in subsequent images. After applying the augmentation strategy, we observed a significant improvement in the quality of the generated images.



Figure 9. Ablation study of the Gaussian noise augmentation.

A.3. Evaluation of Pose Control

In the field of image animation, the mean keypoint distance (AKD) and the average pose distance (APD) are commonly used to evaluate the performance of pose control. We calculated the APD metric of Playmate on two datasets (HDTF and our dataset), as shown in the table below.

Table 5. Evaluation of Pose Control.

Dataset	APD_{jaw}	APD_{pitch}	APD_{roll}
HDTF	3.003°	1.308°	1.214°
Collected dataset	3.714°	1.751°	1.398°

A.4. Disentanglement of Facial Representation

Figure 10 shows the effective disentanglement of the latents of the face. From top to bottom: the raw generated sequence, applying generated poses with fixed initial facial dynamics, and applying generated facial dynamics with fixed initial head pose and predefined spinning poses.



Figure 10. Disentanglement between head pose and facial dynamics.