EP-Talk: Emotion-Aware Personalizable Talking Head Generation

Haopan Ren, Shisheng Huang, Wei Duan, Deqi Li, Wanyu Li, Hua Huang, *Senior Member, IEEE*

Abstract—Despite the remarkable progress in diffusion-based talking head generation, it remains challenging to accurately model stylized facial expressions and personalized head movements. To alleviate these issues, we introduce EP-Talk, an emotion-aware personalizable talking head generation method. Our core idea is to learn *precise* facial emotion and *personalizable* head movement embedding from a short reference video, thereby enhancing the generative model's capacity for stylized feature expression. Specifically, we first propose a diffusion-based natural talking head generation module, which accurately maps audio features to facial dynamics, producing audio-synchronized natural talking videos through a lip-aware displacement predictor module. Then the proposed emotion-aware facial dynamics module and motion-aware personalization module extract relevant features from reference videos, which are leveraged to adjust facial emotions and control head movements, respectively. To demonstrate the effectiveness of our EP-Talk, we conducted extensive self-reenactment and cross-identity reenactment experiments on three datasets: MEAD, HDTF, and CelebV-HQ. The results show that our EP-Talk surpasses all existing approaches in terms of facial emotion accuracy and the personalization of head movements, establishing a new benchmark for stylized expression in the field of audio-driven talking head generation.

Index Terms—Natural Talking Head Generation, Motion-Aware Personalizable Dynamics, Emotion-Aware Facial Dynamics, Feature Alignment, Implicit Dynamic Space.

1 Introduction

With the rapid advancement of video generation technologies, audio-driven talking head generation [1], [2], [3], [4], [5] has become a pivotal research direction in computer vision and computer graphics. It has been widely applied in various tasks such as film production, teleconferencing, online education, and human–computer interaction [6], [7], [8], [9]. By training on large-scale video datasets, the generative model effectively learns the underlying relationships between audio features and head movements, enabling the synthesis of high-quality, long-duration talking head videos. Nevertheless, existing methods frequently overlook the accurate capture of individualized stylistic features [10], [11], [12], particularly facial expressions and personalized head movements, resulting in videos that lack the desired realism.

Audio-driven talking head generation methods [13], [14], [15] typically utilize 3D Morphable Models (3DMMs [16], [17], [18], [19]) or facial landmarks [20], [21] as intermediate representations, where neural networks are employed to learn the mapping between audio features and facial regions in the implicit space. Other methods use textual descriptions or audio-decoupled features as control signals, aligning audio features with facial attributes in the implicit feature space of diffusion models [12], [22], [23] via carefully designed loss functions. A major challenge faced by these methods [24], [25] is the limited effectiveness of audio control signals, which prevents the accurate alignment of audio features with style features in the latent space, leading to instability in the stylized features of the generated videos.

On the other hand, video-driven facial animation approaches align implicit features in the latent space to achieve more precise facial animation. Methods including [26] and [27] extract motion coefficients from videos, while [10], [28], and [29] estimate motion parameters to control facial dynamics via neural texture prediction in tri-plane representations, thereby achieving realistic head animation. Recent work, such as Takin-ADA [30], incorporates emotional signals into the diffusion model to generate emotion-conditioned videos, then adopts a video-driven approach for aligning head motion features in an implicit keypoint space, ultimately enabling emotionally expressive audio-driven facial generation. However, naively applying this video-driven feature alignment strategy to audio-driven talking head generation can easily lead to issues such as poor audio-lip synchronization [31], inaccurate control of facial emotions [32], and limited reconstruction of stylized head motion features [33].

These challenges motivated the development of a novel approach, EP-Talk, aimed at improving stylized feature representation for individual-specific attributes. Our core hypothesis is that pre-trained generative models [12], [22], [23] have learned rich prior knowledge from large-scale video datasets. By leveraging a small set of reference videos from specific individuals, we use a feature alignment strategy in the latent space to enhance facial emotion control [11], [34] and personalized head movement capabilities [35], [36], therefore enhancing the model's ability to generate stylized expressions.

Specifically, the proposed EP-Talk method comprises

- *Haopan Ren, Wei Duan, Deqi Li, and Wanyu Li are with the School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China. E-mail: 202231081023@mail.bnu.edu.cn, 202321081023@mail.bnu.edu.cn, dqli@mail.bnu.edu.cn, 202321081027@mail.bnu.edu.cn.*
- *Shisheng Huang and Hua Huang are with the School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China. E-mail: {huangss, huahuang}@bnu.edu.cn.*

Hua Huang is the corresponding author.

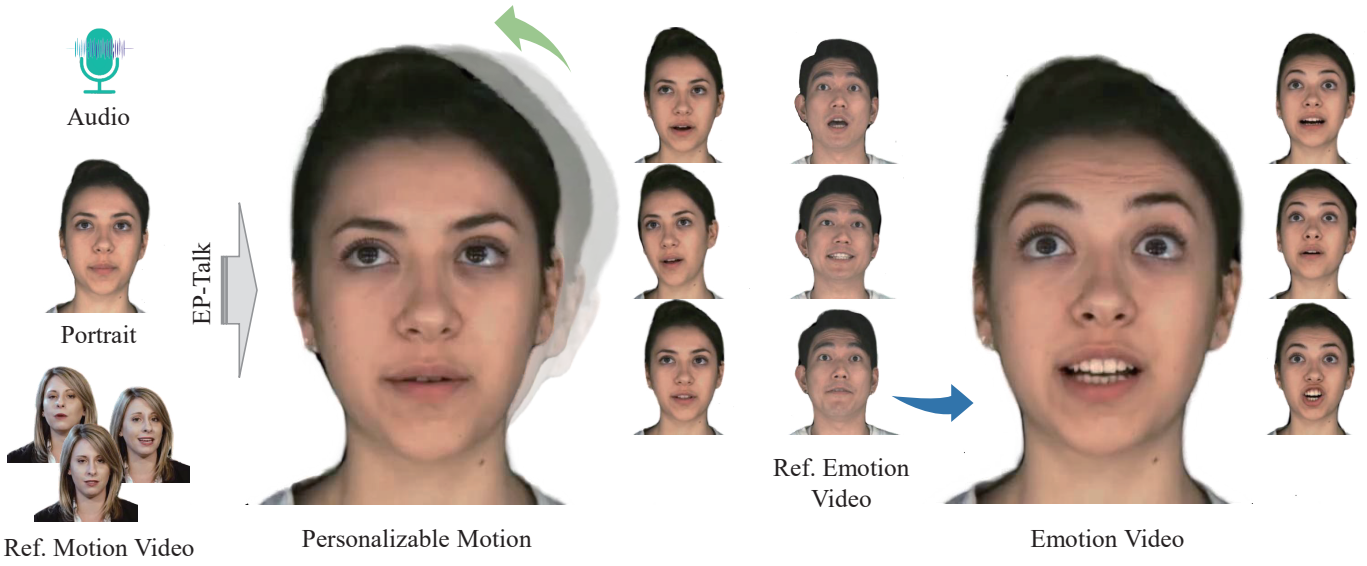

Fig. 1. We propose a novel emotion-aware and personalized talking head generation method, termed EP-Talk. Given an input audio clip, a portrait image, and a reference motion video, EP-Talk generates personalized head-motion videos while ensuring accurate synchronization between speech and lip movements. Likewise, when provided with a reference video depicting facial expressions, our method enables precise control over facial emotions. Benefiting from its emotion-aware personalized feature learning mechanism, EP-Talk preserves richer identity-specific characteristics in both self-reenactment and cross-identity reenactment tasks, particularly in terms of fine-grained emotional expression and personalized head motion.

three key modules: (1) The Natural Talking Head Generation Module, which includes the Audio2Implicit-Dynamics [37] and Implicit2Facial [29] modules, transfers audio features to facial dynamics and employs a lip-aware displacement predictor module to enhance audio-lip synchronization; (2) The Emotion-Aware Facial Dynamics Module, which employs an emotion feature extractor [10], [38] trained on large-scale datasets to accurately extract facial emotion dynamics from the reference video; and (3) The Motion-Aware Personalized Dynamics Module, which uses a personalized motion feature extractor [38], [39] trained on large datasets to extract individual-specific head movement information from the reference video. The extracted motion information is input into the implicit motion-aware displacement predictor to estimate keypoint position changes in the latent space. By employing implicit feature alignment strategies, we achieve personalized head movement manipulation. Finally, the natural talking motion, facial emotion, and personalized head movement features are aligned in the latent space. This multimodal feature alignment enables the generation of audio-driven talking videos with accurate facial emotions and personalized head movements, facilitated by a deformation network and decoder [29].

To validate the effectiveness of our method, we conduct extensive cross-identity reenactment and self-reenactment experiments on three publicly available datasets: MEAD [40], HDTF [41], and CelebV-HQ [42]. These experiments involve comparisons with several state-of-the-art (SOTA) approaches, including SadTalker [31], Diffused Heads [12], Echomimic [23], AniPortrait [20], Hallo2 [37], EAMM [10], EAT [32], IP_LAP [43], and Hallo3 [33]. Performance evaluations primarily focus on facial emotion manipulation and head movement control. Extensive experimental results demonstrate that our EP-Talk consistently outperforms these methods, particularly in preserving the stylized characteristics of target individuals, with notable improvements in the precise control of facial emotions and personalized head movements. We summarize the key contributions as follows:

1) We propose a foundational module for natural talking head generation that enables the synthesis of natural speaking videos with synchronized lip movements.
2) An emotion-aware facial dynamics module is designed to extract emotional features from reference videos for the precise manipulation of facial expressions.
3) A motion-aware personalized dynamics module is introduced to separate motion features from reference videos, guiding the generation of personalized head movements in the synthesized video.

2 Related Work

2.1 Audio-driven Talking Head Video Generation

Early audio-driven talking head generation methods primarily focused on the synchronization between audio and lip movements [44], [45], [46], [47], [48], typically using explicit 3D morphable models (3DMM) or facial landmarks to represent lip motion. Li et al. [16] and Blanz et al. [49] introduced a novel texture-based 3D facial modeling technique that transforms facial shape and texture into spatial vector representations. This approach enables the creation of deformable face models, which allow for the generation of new appearances and expressions through linear combinations. VOCA [13] and MeshTalk [15] introduced audio-driven facial animation frameworks that take arbitrary speech signals as input while enabling flexible control over facial identity, speaking style, head pose, and other attributes.

The widespread application of diffusion models [50], [51], [52], [53], [54], [55] in the field of video generation

has significantly enhanced the quality of image synthesis. AniPortrait [20] introduces a framework that converts audio into 3D motion, maps it to 2D facial landmarks, and uses a diffusion-based generator to create photorealistic facial animations. To address the instability caused by weak audio-driven signals and unnatural results from keypoint-driven methods, Echomimic [23] proposes a novel approach that concurrently trains on both audio and facial landmarks. Hallo2 [37] introduces a patch-drop technique with Gaussian noise to improve visual consistency and temporal coherence in long-duration videos. It also uses a codebook and temporal alignment to achieve 4K resolution synthesis. Additionally, by incorporating semantic textual labels for facial expressions, the method enhances the diversity of the generated content. Although these methods have achieved significant improvements in image quality, long-duration generation, and audio-lip synchronization, they are limited in their ability to capture stylized expressions for specific individuals, particularly in terms of accurate facial emotion manipulation and personalized head movement.

2.2 Control of Facial Emotions

Emotion control [56], [57], [58], [59], [60], [61] signals can be broadly classified into text labels, driving audio, and reference videos, depending on the source of the facial emotion. EAMM [10] introduces an emotion-aware motion model that generates emotional talking head videos by incorporating an emotion source video. EAT [32] introduces an emotion-aware, audio-driven talking head generation approach that achieves efficient emotion control through parameter adaptation. StyleTalk [35] proposes a style-controllable facial generation framework. With the rapid rise of diffusion models, diffusion-based methods for facial emotion control have gradually become mainstream. To enhance identity consistency across video sequences, Hallo3 [33] adopts a pretrained, transformer-based video generation framework. It explores various speech audio conditioning and motion frame mechanisms to enable audio-driven emotional video generation.

The use of text labels [62], [63], [64], [65] can only represent facial emotions in a discrete, coarse-grained manner, making accurate facial emotion control difficult to achieve. Some methods [66], [67], [68], [69], [70] that decouple facial emotions from audio in video generation often result in ambiguity. Although using additional reference videos can control facial features, generative models trained on large-scale video datasets learn complex and diverse emotional expressions, which limits their ability to accurately manipulate facial emotions specific to individual identities.

2.3 Control of Head Movements

In talking head video generation, head pose [12], [21], [71], [72], [73], [74], [75] plays a crucial role, directly influencing the dynamic effects of the generated video. Current methods primarily rely on audio signals or predefined poses to control head movements. Early explicit 3DMM methods employed template matching algorithms to align the head model with 3D scan data, representing head motion through a set of feature vectors. However, such methods often result in lower image quality. Recognizing the critical role of head pose in talking-head video generation, EAMM [10] and EAT [32] employ off-the-shelf head pose estimators to extract per-frame head pose coefficients from training videos. PDFGC [38] introduces a novel head synthesis approach, employing a progressive decoupled representation learning strategy. This strategy disentangles factors such as identity, head pose, eye gaze direction, emotion, and lip movements from coarse to fine levels.

Although current methods [76], [77], [78] achieve coordination between head pose and audio rhythm, weak audio control signals result in ambiguous head movements. On the other hand, methods that extract head pose sequences from reference videos provide stronger control signals but lack personalized head movement features specific to individual identities.

Inspired by previous approaches, we propose EP-Talk, which leverages a small amount of video data to extract precise control signals for facial emotions and head movements. By aligning multimodal features in the latent space, the proposed method enhances the expressive capacity of pretrained generative models.

3 METHOD

The audio-driven talking head generation method generally extracts control signals from the input audio a_t and identity features from a reference portrait I_s. It then employs neural networks to model the mapping relationship between audio features and facial animations. Finally, high-fidelity talking videos are generated through rendering techniques.

As shown in Fig. 2, the EP-Talk framework consists of three key modules: Natural Talking Head Generation (NTHG), Emotion-Aware Facial Dynamics (EAFD), and Motion-Aware Personalized Dynamics (MAPD). Firstly, in the NTHG module, the Audio2Implicit-Dynamics (A2ID) module generates a talking video from the input portrait I_s and audio a_t, through denoising and noise addition. The Lip-Aware Displacement Predictor (LADP) predicts lip movement deviations based on audio features and identity coefficients. The Implicit2Facial (I2F) module refines the video by aligning these deviations and generating a natural talking video V_n using a deformation network W and decoder D. Secondly, in the EAFD module, the emotion feature extractor E_e extracts emotional features from the reference video V_e, which are processed by the Implicit Emotion-Aware Displacement Predictor (IEDP, P_e) to predict emotional feature changes. Thirdly, in the MAPD module, personalized head motion features are extracted from the reference video V_m using a head movement extractor E_{pm}, which are then processed by the Implicit Motion-Aware Displacement Predictor (IMDP, P_m) to predict head movement deviations. Finally, the emotional and head movement features are aligned in the implicit space, and the I2F module generates a stylized talking video V_f.

3.1 Natural Talking Head Generation

The NTHG module consists of three submodules: A2ID, I2F, and LADP. The A2ID module is composed of a diffusion model that generates a coarse speaking video based on the driving audio and reference portrait. The I2F module,

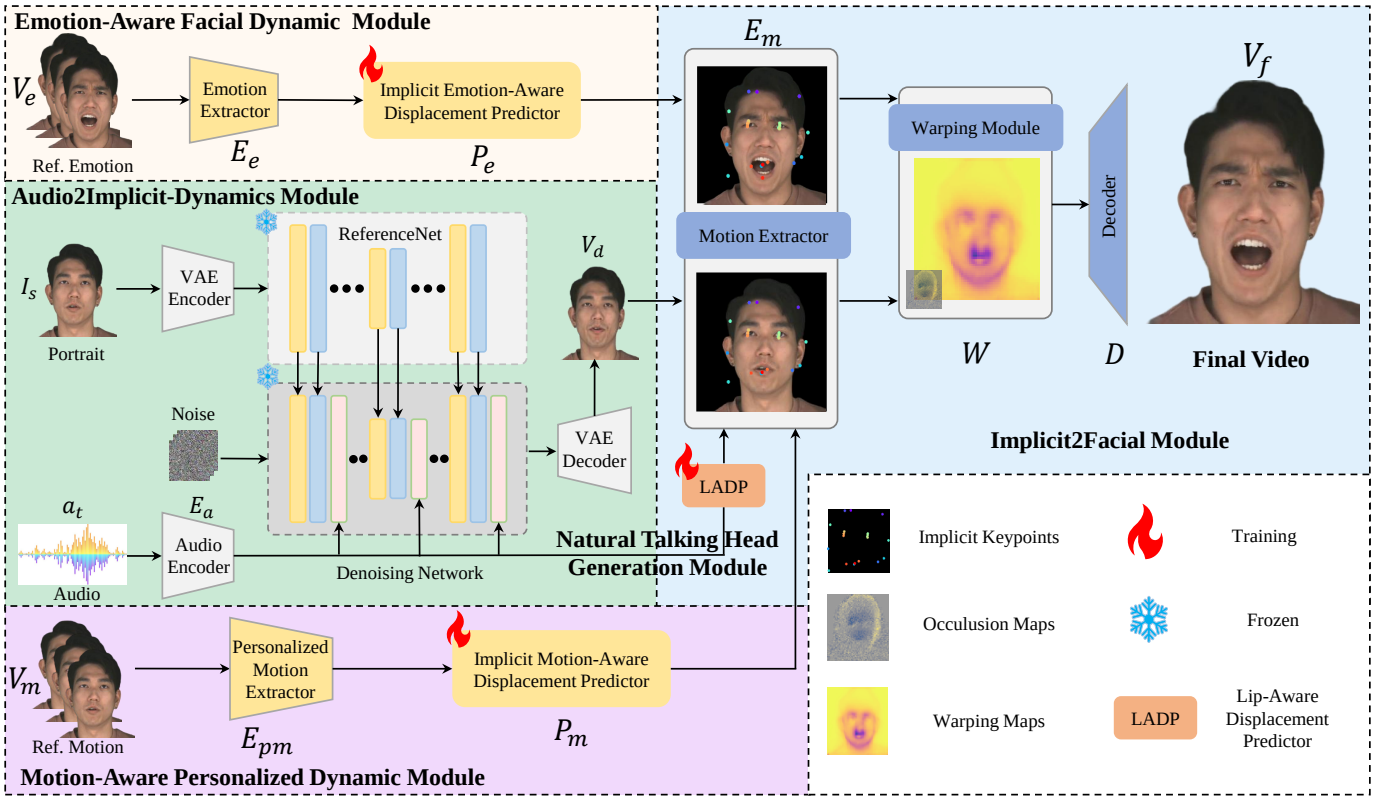

Fig. 2. The pipeline of our EP-Talk framework. The framework consists of four key stages: (i) In the first stage, the A2ID module generates a coarse talking-head video via a diffusion process conditioned on the input audio and the reference portrait. (ii) In the second stage, the I2F module synthesizes a natural and identity-consistent talking-head video based on the coarse video and the reference portrait. Leveraging the extracted audio features and identity coefficients, the LADP module predicts the deviation of mouth-related features in the latent space, aiming to further enhance the synchronization between speech and lip movements. (iii) In the third stage, the EAFD module extracts fine-grained facial emotion features from the reference emotion video and predicts the displacement of emotion features in the latent space. (iv) In the final stage, personalized head-motion features extracted by the MAPD module are fed into the IMDP module, which predicts head-motion deviations in the latent space. By aligning multimodal features within the latent space, EP-Talk ultimately generates a talking-head video with precise emotional expressions and personalized head movements.

which utilizes an implicit keypoint model, generates high-quality video outputs in a video-driven manner. Simply concatenating the two modules may result in inaccurate lip movements in the mouth region. Therefore, the LADP module is proposed to improve synchronization between the audio and lip movements.

Audio2Implicit-Dynamics Module. The audio-driven portrait generation is modeled as a denoising diffusion process of the portrait image conditioned on audio features. Specifically, given an input portrait image I_s and audio clip a_t, we encode I_s as a latent vector z_0 using a VAE encoder, and a_t to audio feature vector a_e. Then we map z_0 to Gaussian noise vector z_T following T steps of additive Gaussian noise, i.e., $\{z_t\}_{t=1}^{T}$ following:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

where $t \in \{1, 2, \ldots, T\}$ denotes the diffusion steps, $\alpha_t = 1 - \beta_t$ with $\beta_t \in (0, 1)$ being the variance schedule, and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ is the cumulative product of α_t. Then we formulate a audio conditioned noise prediction $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(z_t, t, a_e)$ and reconstruct the portrait latent vector z_0 via the reverse diffusion process following:

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(z_t, t, a_e)\right) + \sigma_t n, \quad n \sim \mathcal{N}(0, I). \tag{2}$$

The final latent vector z_0 is decoded to an feature embeddings V_d using a VAE decoder, and mapped to implicit keypoints x_d using the motion extractor E_m as $x_d = E_m(V_d)$.

Implicit2Facial Module. Once the implicit keypoints are extracted, an I2F block is employed to map the keypoints to the final portrait image within a video-driven framework. Specifically, the I2F module integrates two components including a warping network W and a image decoder D. The warping network W maps the appearance feature map $f_s = \mathcal{F}(I_s)$ of portrait image I_s and implicit keypoints x_d to a dense warping field $W(f_s, x_d)$, and then fed to the image decoder D to generate high-quality video output, as formally expressed below:

$$x_s = s_s * (x_{c,s} R_s + \delta_s) + t_s, \tag{3}$$

$$x_d = s_d * (x_{c,s} R_d + \delta_d) + t_d, \tag{4}$$

$$V_f = D(W(f_s, x_d, x_s)), \tag{5}$$

where the R represents the rotation matrix, t denotes the translation matrix, δ refers to the facial emotion, s represents the scale, $x_{c,s}$ denotes the implicit keypoint positions of the source image, and x_d represents the implicit keypoint positions of the driving image. f_s denotes the extracted appearance features and V_f represents the final target talking video.

Lip-Aware Displacement Predictor. The A2ID and I2F

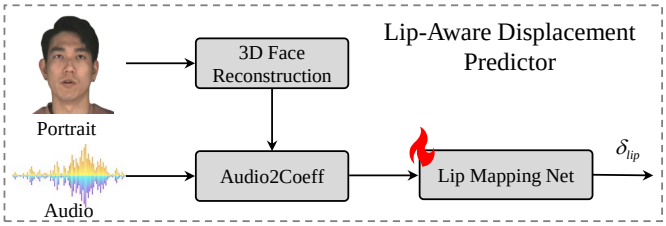

Fig. 3. The architecture of the lip-aware displacement predictor.

modules have successfully achieved high-quality video generation. However, when combined, they still exhibit notable issues with audio-lip synchronization. To address this, the LADP module is proposed, which learns lip movement features from audio and leverages video supervision to improve lip synchronization accuracy.

As shown in Fig. 3, the LADP module primarily consists of three components: the 3D facial reconstruction module, the Audio2Coeff module, and the lip mapping network. Identity coefficients are extracted from the reference portrait, and audio features are then used to drive the process, with the mapping network producing implicit keypoint deviations (δ_{lip}) of the lip features. The natural talking video V_n with synchronized lip movements is subsequently generated through the application of Equation 5.

3.2 Emotion-Aware Facial Dynamics

The emotional control signals extracted from audio or text are weak, leading to instability or inaccuracy in facial emotion control. Therefore, we choose to extract emotional features from reference videos of specific individuals to manipulate the synthesis of facial emotions. Specifically, a pre-trained emotional feature extractor E_e is employed to derive features from the reference video V_e, encompassing attributes such as the individual's identity, lip movements, facial expressions, and head poses. To accurately capture the emotional features of the face and eliminate interference from other attributes, only the facial emotional features f_{exp} are selectively extracted and input into the IEDP module to predict emotional changes. The emotion feature prediction process is formally represented as follows:

$$f_{exp} = E_e(V_e), \quad \delta_{exp} = P_e(f_{exp}), \tag{6}$$

where the E_{e} represents the feature encoder, V_{e} represents the reference emotional video. f_{exp} refers to the extracted emotion feature vectors. P_{e} represents the Implicit Emotion-Aware Displacement Predictor, and δ_{exp} indicates the predicted deviation of the emotional features. The comprehensive structure of the EAFD module is provided in the supplementary materials.

The EAFD module maps the extracted emotional features into the implicit space, where they are fused with the implicit features generated from the NTHG stage. Finally, a deformation network W and decoder D are employed to generate the talking video with facial emotions. This process can be formally expressed as:

$$\delta_d^{'} = \delta_d + \delta_{exp}, \tag{7}$$

$$x_d^{'} = s_d * (x_{c,s} R_d + \delta_d^{'}) + t_d, \tag{8}$$

$$V_{oe} = D(W(f_s, x_d^{'}, x_s)), \tag{9}$$

where $x_d^{'}$ represents the 3D spatial keypoint positions corresponding to the implicit facial expression features. V_{oe} represents the generated talking video with accurate facial emotions. Since facial expressions only affect the facial region, the implicit keypoint sequence is selected exclusively for the facial area. When incorporating facial emotion features, it is essential to maintain the alignment of keypoints between the NTHG phase and the facial emotion keypoints in the implicit space.

3.3 Motion-Aware Personalized Dynamics

To accurately represent the personalized head motion characteristics of specific individuals in reference videos, two types of control parameters are employed to extract head dynamics. The decoupling method [38] is trained on a large dataset of 2D images, enabling precise control over various aspects such as facial expressions, gaze direction, mouth movements, and head poses. On the other hand, the registration method [16] aligns 3D scanned data and uses template matching algorithms to accurately match the head pose in space, thereby facilitating precise head motion control.

Initially, the encoder E_{pm} is utilized to extract personalized motion features from the reference video V_m. The motion parameters f_{pose} are then fed into the IMDP module P_m to estimate the deviations R_{pose} in head movement. The process of personalizing head motion can be formally represented by the following equation:

$$f_{pose} = E_{pm}(V_m), \quad R_{pose} = P_m(f_{pose}), \tag{10}$$

where V_m denotes the motion reference video, E_{pm} is the motion feature extractor, f_{pose} represents the extracted pose parameters, P_m is the implicit motion-aware displacement predictor module, and R_{pose} represents the predicted pose variation deviation.

Similarly, the MAPD module maps the extracted personalized head motion features into an implicit space, which are then fused with the implicit features generated from the NTHG stage. Finally, a deformation network W and decoder D are employed to generate a talking video V_{om} with stylized head motion specific to the individual. This process is formally expressed as:

$$R_d^{'} = R_d + R_{pose}, \tag{11}$$

$$x_d^{''} = s_d * (x_{c,s} R_d^{'} + \delta_d) + t_d, \tag{12}$$

$$V_{om} = D(W(f_s, x_d^{''}, x_s)), \tag{13}$$

where the term $R_d^{'}$ denotes the pose parameters associated with personalized head movements, while $x_d^{''}$ represents the 3D spatial keypoint positions corresponding to the implicit stylized head motion features. V_{om} represents the generated talking video with personalized head movements.

3.4 Training Optimization

Due to the limited amount of training data, using traditional loss functions such as identity loss, mean squared error loss, and generative adversarial loss would severely degrade the quality of the generated images. Therefore, in the experiments, a feature extractor trained on a large-scale

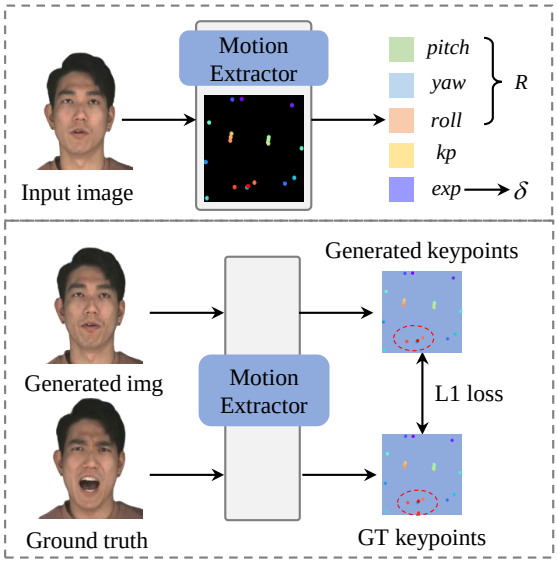

Fig. 4. The architecture of the feature extractor and the loss settings.

dataset is employed to compute the keypoint displacement loss in the implicit feature space. As shown in the Fig. 4, the motion extractor can extract head rotation features, standard keypoint features, and facial expression features from the image. This process can be formally expressed as follows:

$$R_p, R_y, R_r, x_{kp}, \delta = E_m(I), \tag{14}$$

where R_p, R_y, and R_r refer to the pitch angle, yaw angle, and roll angle, respectively. x_{kp} represents the standard keypoint positions extracted from the image, while δ denotes the facial emotion keypoint positions extracted from the image.

To accurately train for different types of attributes, a pre-trained feature extractor is employed to extract implicit keypoint locations. Distinct implicit keypoint sequences are defined for different regions, such as the lips, face, and head, with an L1 loss function applied to constrain variations in these keypoints. Through this decoupling strategy, our approach effectively isolates unrelated features, preventing the degradation of synthesized video quality caused by feature coupling.

4 Experiments

4.1 Experiment Settings

Implementation Details. During training, the initial learning rate for all three training modules is set to 2×10^{-4}, with each module sharing similar network architectures. In the A2ID module, feature maps with a resolution of 512×512 are initially generated, followed by downsampling to reduce the resolution. The motion extractor is then applied to obtain 21×3 implicit keypoint control displacements. In the I2F module, implicit keypoints and the reference portrait are used as inputs, and high-quality video sequences are generated through a warping network and a decoder, with the synthesized video resolution being 512×512. To ensure high-quality lip synchronization, we employ the pre-trained feature extractor from SadTalker. For emotion feature extraction, the method proposed in [38] is employed to extract facial expression parameters. For personalized head pose control, two approaches [13], [38] are utilized to extract head motion coefficients.

Datasets. The effectiveness of the proposed EP-Talk is evaluated on three datasets: **MEAD [40]**, **HDTF [41]**, and **CelebV-HQ [42]**, with a primary focus on facial emotion control and personalized head motion. The MEAD dataset is a large-scale multimodal resource designed for studies on facial emotion expression and audio-visual synchronization, featuring comprehensive emotional annotations and diverse facial actions. The HDTF dataset, sourced from YouTube, includes videos with varying resolutions and recording conditions and is widely used in research on head motion and facial expression variations. The CelebV-HQ dataset is a high-resolution collection of celebrity portrait videos, encompassing a wide range of identities, expressions, head poses, lighting conditions, and scene variations.

Evaluation Metrics. A range of evaluation metrics is employed to comprehensively assess the model's performance, covering various aspects such as the quality of generated images, audio-lip synchronization, facial emotion representation, and head motion amplitude. These metrics include Fréchet Inception Distance (FID), Fréchet Video Distance (FVD), Synchronization-C (Sync-C), Synchronization-D (Sync-D), and E-FID. Specifically, FID and FVD [79] measure the similarity between generated and real images at the image and video frame levels, respectively, with lower values indicating better performance. The SyncNet [80] scores evaluate the quality of lip synchronization, where higher Sync-C scores and lower Sync-D scores indicate better alignment with the driven audio signal. To further evaluate the expressiveness of facial emotions in the generated videos, E-FID [81] is employed, which quantitatively measures the divergence in facial expressions between the synthesized and ground-truth videos. Additionally, the Face Landmark Distance (F-LMD) and the Average Expression Distance (AED) are employed to quantify the disparity in facial emotions, the Average Pose Distance (APD) [82] and the pose factor of the PDFGC method are utilized to assess variations in head poses, and the Cosine Similarity of Identity Embeddings (CSIM) is used to evaluate identity preservation.

4.2 Experimental Results on the MEAD Dataset

As demonstrated in Tab. 1, cross-identity emotion-driven synthesis is performed on the MEAD dataset, comparing the proposed method with SOTA approaches in terms of video quality, synchronization, and emotion control. In terms of image quality, the proposed EP-Talk method achieves the best results, with an FID of 17.78 and an FVD of 55.31. Compared to other methods, the proposed approach demonstrates a significant improvement in image quality. In terms of audio and lip synchronization, the proposed EP-Talk method achieves results comparable to Hallo2, with scores of 10.67 and 10.54, respectively. In terms of facial emotion, our method leverages video to provide accurate facial emotion signals, significantly outperforming other methods in facial expression intensity. The EAT method, which also uses the same video reference approach, yields comparable experimental results (0.377 vs. 0.380). Based on these experimental results, it can be concluded that the facial

Fig. 5. Comparison of qualitative experimental results on the MEAD dataset. The yellow box marks identity preservation issues, the red box highlights texture artifacts in mouth shapes, and the blue box indicates mismatches between the generated facial expression and the reference video.

TABLE 1
A comparison of cross-identity reenactment performance on the MEAD and HDTF datasets. The blue and green areas represent the first and second rankings, respectively. Downward arrows indicate that lower values are preferred, while upward arrows indicate that higher values are preferred.

Method	MEAD					HDTF				
	FID ↓	FVD ↓	Sync-C ↑	Sync-D ↓	E-FID ↓	FID ↓	FVD ↓	Sync-C ↑	Sync-D ↓	E-FID ↓
SadTalker [31]	21.67	59.12	0.688	12.31	0.472	56.38	50.57	0.309	13.61	0.172
Diff heads [12]	29.02	118.0	1.824	12.52	1.033	216.9	166.0	0.973	12.52	1.286
EchoMimic [23]	21.48	56.03	1.263	11.21	0.438	66.34	49.64	0.353	10.97	0.133
AniPortrait [20]	21.55	55.70	1.046	11.54	0.482	68.44	55.28	0.223	13.10	0.255
Hallo2 [37]	22.38	61.02	1.767	10.54	0.654	40.63	55.42	2.569	10.30	0.227
EAMM [10]	20.07	59.78	0.828	12.52	0.662	77.22	79.62	2.003	12.95	0.380
EAT [32]	20.50	56.32	1.993	10.85	0.380	61.30	53.52	0.976	11.60	0.321
IP_LAP [43]	18.81	65.05	1.002	11.46	0.449	58.93	50.68	0.790	11.56	0.133
Hallo3 [33]	18.28	79.83	1.241	11.65	0.433	45.35	56.02	1.591	11.69	0.196
EP-Talk	17.78	55.31	2.003	10.67	0.377	40.35	45.64	2.946	10.10	0.120

emotion features derived from the video are more accurate than those provided by other emotion control methods.

The qualitative experimental results are shown in Fig. 5. In the Diffused Heads and EAMM methods, the yellow dashed box regions exhibit noticeable blurring and artifacts, making it difficult to preserve the person's identity. While the SadTalker method produces relatively clear facial contours, the facial features are overly smoothed, lacking the detailed characteristics of the person in the reference portrait. The AniPortrait, EchoMimic, IP_LAP, and Hallo3 methods show significant blurring of the teeth or lip features in the mouth region. Furthermore, the IP_LAP method requires extracting identity features from a video, whereas other methods extract identity features from a single image. In terms of facial emotion control, our EP-Talk closely aligns with the facial emotions in the reference video, demonstrating accurate control over facial emotion features. Although the EAT method also uses a similar video reference approach, there is a noticeable disparity between its facial emotions and those in the reference angry video.

4.3 Experimental Results on the HDTF Dataset

Considering that the HDTF dataset videos contain certain facial emotion features, facial emotion comparison experi-

Fig. 6. Comparison of qualitative experimental results on the HDTF dataset. The yellow box highlights identity mismatches with the reference portrait, the red box indicates lip-synchronization errors or degraded mouth-texture quality, and the blue box shows inconsistencies between the subject's head pose and the reference motion video.

ments are conducted first, followed by head pose comparison experiments. As shown on the right side of Tab. 1, cross-identity reenactment experiments are performed on the HDTF dataset, comparing the proposed method with the latest approaches in terms of video quality, audio-lip synchronization, and facial expression control. In terms of image and video generation quality, the proposed EP-Talk method achieved the best experimental results (FID: 40.35, FVD: 45.64). For speech and lip synchronization, EP-Talk also delivered the best performance, demonstrating competitive synchronization capabilities when compared to Hallo2 (Sync-D: 10.10 vs. 10.30). The proposed EP-Talk method shows a clear advantage in facial emotion control, outperforming all other methods. Since the facial emotions in the HDTF dataset are not as prominent, the EchoMimic and IP_LAP methods achieved results comparable to EP-Talk (0.133 vs. 0.120).

The qualitative experimental results are shown in Fig. 6. The images generated by the SadTalker, DiffusedHeads, and EAMM methods exhibit low quality and fail to maintain consistent identity in terms of the person's appearance. The Hallo2 and AniPortrait methods suffer from noticeable blurring or misalignment in the mouth region, as indicated by the red dashed boxes. Although the EchoMimic, IP_LAP, and Hallo3 methods achieve relatively better results, the generated speaking videos lack personalized head motion features, showing a significant discrepancy from the reference motion videos, as highlighted by the blue boxes in Fig. 6. The proposed EP-Talk exhibits distinct advantages in image generation quality, audio-lip synchronization, and head motion control.

TABLE 2
Comparison of quantitative results for expression control on the CelebV-HQ dataset. All compared methods perform cross-identity reenactment by using identity images from the CelebV-HQ dataset and expression references from the MEAD dataset.

Method	AED ↓	E-FID ↓	F-LMD ↓	CSIM ↑
SadTalker [31]	0.157	0.071	0.063	0.552
Diff heads [12]	0.173	0.295	0.131	0.030
EchoMimic [23]	0.208	0.075	0.058	0.678
AniPortrait [20]	0.207	0.074	0.081	0.901
Hallo2 [37]	0.196	0.088	0.078	0.872
EAMM [10]	0.166	0.097	0.053	0.499
EAT [32]	0.171	0.074	0.053	0.514
IP_LAP [43]	0.155	0.089	0.081	0.914
Hallo3 [33]	0.186	0.093	0.082	0.798
EP-Talk	0.124	0.070	0.046	0.911

4.4 Experimental Results on the CelebV-HQ Dataset

As shown in Tab. 2, emotion videos and driving audio from the MEAD dataset, along with identity portrait images from the CelebV-HQ dataset, are selected to compare cross-identity driving results. This section primarily focuses on facial emotion control and identity preservation. The proposed method achieves the best results in facial emotion control, while identity preservation performance is comparable to that of the IP_LAP method. Notably, the IP_LAP method requires reference portrait videos as input to effectively maintain identity features.

As shown in Tab. 3, motion videos and driving audio from the HDTF dataset, along with identity portrait images

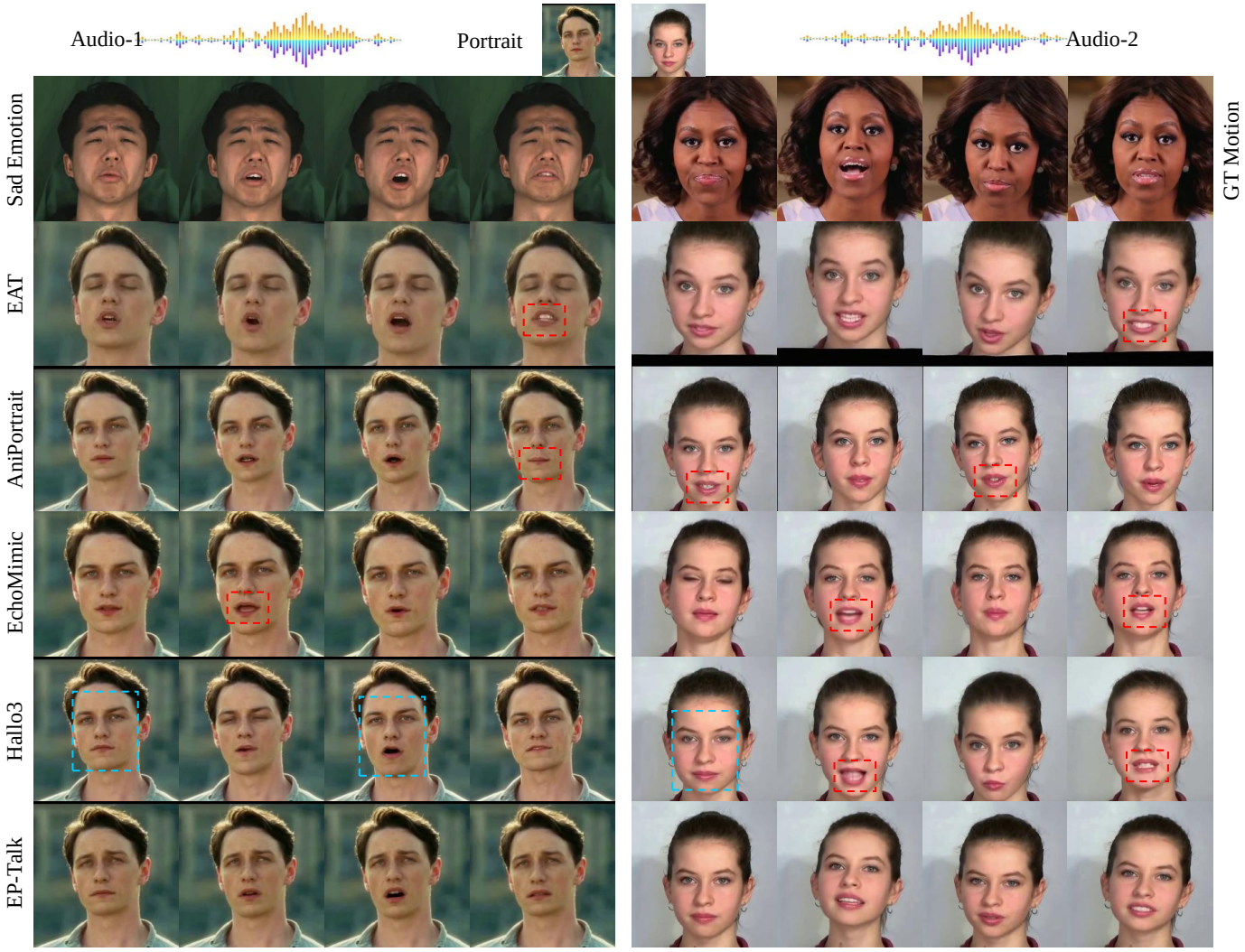

Fig. 7. Comparison of qualitative experimental results on the CelebV-HQ dataset. The left part of the figure demonstrates expression-driven talking-head generation using the reference emotion video, portrait image, and driving audio. The right part shows motion-driven talking-head generation using the motion reference video, portrait image, and driving audio. The red box indicates lip-synchronization errors or degraded mouth-texture quality, while the blue box highlights failures to reproduce the intended expression or motion.

TABLE 3
Quantitative comparison of motion-generation performance on the CelebV-HQ dataset. All evaluated methods use identity images from CelebV-HQ and motion references from the HDTF dataset for cross-identity reenactment.

Method	APD ↓	F-LMD ↓	PDFGC ↓	CSIM ↑
SadTalker [31]	0.540	0.087	0.741	0.624
Diff heads [12]	1.435	0.106	1.518	0.245
EchoMimic [23]	0.601	0.120	1.271	0.600
AniPortrait [20]	0.459	0.107	0.574	0.752
Hallo2 [37]	0.351	0.089	0.471	0.801
EAMM [10]	0.784	0.130	2.383	0.569
EAT [32]	0.899	0.094	2.162	0.498
IP_LAP [43]	0.386	0.121	0.480	0.834
Hallo3 [33]	0.463	0.115	0.585	0.747
EP-Talk	0.264	0.085	0.437	0.838

from the CelebV-HQ dataset, are selected to compare cross-identity driving results. This section primarily focuses on the ability to preserve personalized head movement and identity in the generated images. The proposed EP-Talk outperforms all other approaches in personalized head pose movement, achieving the best results in terms of APD, F-LMD, and PDFGC, while also demonstrating strong identity preservation (CSIM). The SadTalker, Hallo2, and Hallo3 methods generate videos with corresponding motions synchronized to the audio, leading to superior performance in head movement control.

The left side of Fig. 7 presents the facial expression control results, while the right side illustrates the head motion results. Four methods are primarily selected for cross-identity reenactment comparison. The red boxes in the figure indicate regions where the mouth is blurred or where audio-lip synchronization is misaligned, while the blue boxes highlight inaccuracies in emotion or head motion. In the facial emotion control comparison on the left, the images generated by the EAT, AniPortrait, and EchoMimic methods exhibit noticeable blurring or inaccurate mouth movement in the mouth region. Although Hallo3 generates relatively good results, the text descriptions fail to provide accurate facial emotion control signals, resulting in a lack of

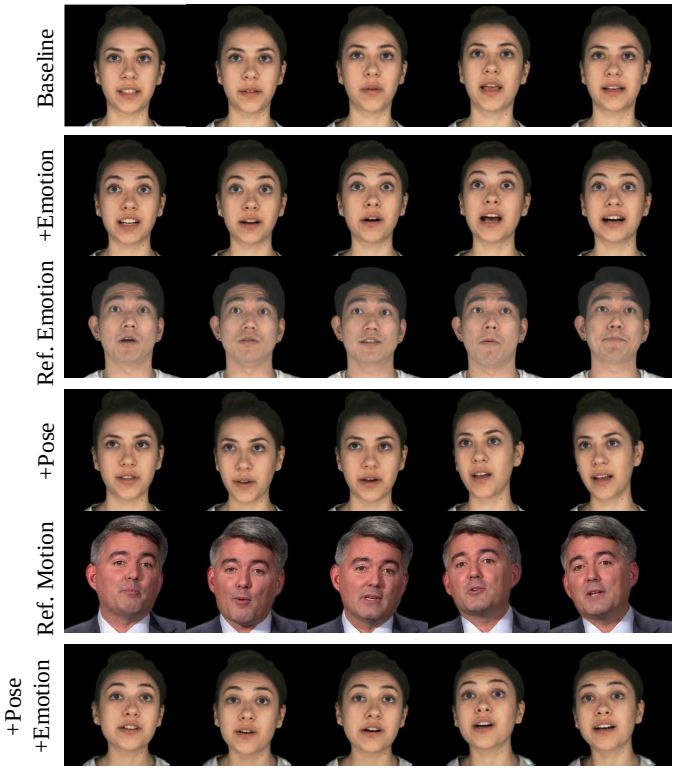

Fig. 8. Experimental results of joint control of facial emotions and head motion.

the sad expression present in the reference emotion video. In contrast, EP-Talk leverages emotion videos as references, enabling the synthesis of talking videos with more accurate facial emotions. In the head motion comparison on the right, the images generated by AniPortrait, EchoMimic, and Hallo3 show poor texture detail in the mouth region. While the EAT method generates corresponding head motions, the quality of the generated images is relatively low. The proposed EP-Talk demonstrates clear advantages in both image generation detail and personalized head motion. Consistent with the quantitative experimental results, our method achieves more accurate facial emotion control and stylized head movements.

4.5 Experimental Results of Joint Emotion and Pose Control

To better assess the effectiveness of emotion control and head movement, emotion videos from the MEAD dataset and motion videos from the HDTF dataset are selected as references. The cross-identity reenactment results are shown in Fig. 8. Both the emotion and motion modules can be controlled independently or in combination.

The proposed NTHG module generates natural talking videos based on driving audio and reference portraits, as depicted by the ”baseline” row. The EAFD module is employed to extract emotional features from the reference video and map them into an implicit space to generate talking videos with facial emotions. The ”Ref.Emotion” row represents the reference emotional video, while the ”+Emotion” row illustrates the generated talking video with emotional expressions. Subsequently, the proposed MAPD module extracts personalized head movements from the reference video, aligns the features in the implicit space, and generates talking videos with personalized head movements. The ”Ref.Motion” row represents the reference motion video, while the ”+Pose” row illustrates the generated talking video with stylized head movements. By utilizing the reference facial expressions provided in the ”Ref.Emotion” row and the head movements from the ”Ref.Motion” row, the proposed EP-Talk generates talking videos with accurate facial emotions and personalized head movements, as illustrated in the ”+Pose +Emotion” row.

4.6 Ablation Study

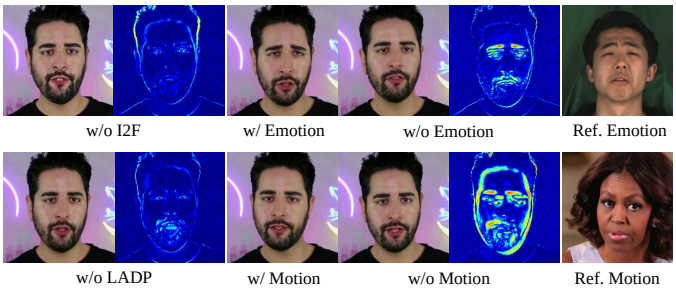

Fig. 9. Qualitative results of the ablation study. In the qualitative ablation study, the contribution of each module is evaluated independently. The blue heat map visualizes the differences in the generated images between the configurations with and without the complete module.

To accurately assess the contribution of each module in the proposed method, six different configuration experiments were conducted, and the results were evaluated in terms of image quality, lip synchronization, emotion control, and head pose. Due to the absence of the I2F module, subsequent modules could not be incorporated, meaning that experiments without I2F essentially use only the A2ID module as the baseline. As shown in Tab. 4, the ”w/o I2F” configuration leads to an overall degradation in model performance. The I2F and LADP modules provide noticeable improvements in image quality and lip synchronization. The ”w/o EAFD” configuration results in a significant reduction in emotion control, while the ”w/o MAPD” configuration leads to a substantial decrease in head movement accuracy. The proposed EP-Talk adopts a balanced strategy, achieving superior experimental results.

The qualitative ablation experiment results are shown in Fig. 9. Given the driving audio, reference portrait, motion video, and emotion video, the outputs of each module are used to generate images and heatmaps. In the ”w/o I2F” configuration, a subtle blur is observed across the entire face. The ”w/o LADP” configuration leads to discrepancies in the lip region. In the ”w/o Emotion” configuration, the left side shows the emotion-driven results, the middle shows the absence of facial emotion, and the right side displays the emotion deviation heatmap. In the ”w/o Motion” configuration, the left side shows the pose-driven results, the middle shows the lack of head pose, and the right side presents the pose deviation heatmap. The last column represents the reference emotion and motion videos. The absence of facial emotion features results in a larger disparity in the facial region of the heatmap, while the lack of head motion features causes a greater difference in the contour region.

TABLE 4
Comparison of ablation study results on selected data from three datasets.

I2F	LADP	EAFD	MAPD	FID ↓	FVD ↓	Sync-C ↑	Sync-D ↓	AED ↓	APD ↓	E-FID ↓	F-LMD ↓
				51.20	48.22	3.511	9.902	0.249	0.793	0.683	0.446
✓		✓	✓	51.87	57.24	4.012	9.047	0.136	0.449	0.612	0.384
✓	✓			30.83	44.74	4.372	9.148	0.226	0.878	0.658	0.489
✓	✓		✓	39.59	52.88	3.446	9.722	0.219	0.407	0.651	0.369
✓	✓	✓		40.34	48.67	4.534	8.548	0.140	0.837	0.585	0.460
✓	✓	✓	✓	41.75	45.98	4.387	9.210	0.137	0.395	0.633	0.366

Consistent with the quantitative experimental results, the proposed EAFD and MAPD modules enable accurate facial emotion control and personalized head motion.

4.7 Limitations

Although EP-Talk successfully achieves facial expression control and personalized head movements, several challenges remain: (1) In the current framework, head motion features are primarily derived from a pre-trained motion extractor to obtain different types of attribute representations. However, due to the inherent performance limitations of this pre-trained model, certain subtle and fine-grained details may not be fully preserved in the synthesized results. (2) While the motion and emotion feature extractors have been trained on large datasets, they still exhibit limited performance when handling large head movements and exaggerated facial expressions. (3) Currently, data for specific individuals is difficult to obtain. However, if a sufficient amount of data can be collected, it would further enhance the model's ability to express stylized features.

5 Conclusion

Aiming to achieve stylized feature representation, we propose an emotion-aware personalizable talking head generation method, EP-Talk. Our core idea is to leverage the prior knowledge of pre-trained generative models and utilize a small amount of video data to provide precise control information, thereby enabling accurate facial emotion control and personalized head movement manipulation within the latent feature space. The proposed NTHG module generates a natural talking video based on the driving audio and reference portrait. To achieve accurate facial emotion control, the EAFD module extracts emotion features from the reference video and enhances these facial emotion features within the latent feature space. The MAPD module extracts motion features from the reference video and enhances personalized head movement features within the latent feature space. Furthermore, the implicit space feature alignment strategy can effectively enhance the ability of pre-trained models to express stylized features. The proposed EP-Talk project aims to advance the controllable and stylized generation of talking avatars.

Future work will focus on two key aspects to enhance the performance of the current model. First, considering the limitations of pre-trained models, incorporating diverse training data, such as exaggerated expressions and intense movements, is expected to improve the model's generalization ability. Second, few-shot training strategies will be explored to enhance the model's representational capacity under limited data.

References

[1] H. X. Pham, S. Cheung, and V. Pavlovic, "Speech-driven 3d facial animation with implicit emotional awareness: A deep learning approach," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 80–88.

[2] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 484–492.

[3] W. Wang, Y. Wang, J. Sun, Q. Liu, J. Liang, and T. Li, "Speech driven talking head generation via attentional landmarks based representation." in *INTERSPEECH*, 2020, pp. 1326–1330.

[4] S. Wang, L. Li, Y. Ding, C. Fan, and X. Yu, "Audio2head: Audio-driven one-shot talking-head generation with natural head motion," *arXiv preprint arXiv:2107.09293*, 2021.

[5] R. Huang, W. Zhong, and G. Li, "Audio-driven talking head generation with transformer and 3d morphable model," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 7035–7039.

[6] L. Sun, "Research on the application of 3d animation special effects in animated films: taking the film avatar as an example," *Scientific Programming*, vol. 2022, no. 1, p. 1928660, 2022.

[7] K. Yu, G. Gorbachev, U. Eck, F. Pankratz, N. Navab, and D. Roth, "Avatars for teleconsultation: Effects of avatar embodiment techniques on user perception in 3d asymmetric telepresence," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 11, pp. 4129–4139, 2021.

[8] A. Alam and A. Mohanty, "Facial analytics or virtual avatars: competencies and design considerations for student-teacher interaction in ai-powered online education for effective classroom engagement," in *International Conference on Communication, Networks and Computing*. Springer, 2022, pp. 252–265.

[9] Q. Cao, H. Yu, P. Charisse, S. Qiao, and B. Stevens, "Is high-fidelity important for human-like virtual avatars in human computer interactions?" *International Journal of Network Dynamics and Intelligence*, vol. 2, no. 1, pp. 15–23, 2023.

[10] X. Ji, H. Zhou, K. Wang, Q. Wu, W. Wu, F. Xu, and X. Cao, "Eamm: One-shot emotional talking face via audio-based emotion-aware motion model," in *ACM SIGGRAPH 2022 conference proceedings*, 2022, pp. 1–10.

[11] C. Zhang, C. Wang, J. Zhang, H. Xu, G. Song, Y. Xie, L. Luo, Y. Tian, X. Guo, and J. Feng, "Dream-talk: Diffusion-based realistic emotional audio-driven method for single image talking face generation," *arXiv preprint arXiv:2312.13578*, 2023.

[12] M. Stypułkowski, K. Vougioukas, S. He, M. Zieba, S. Petridis, and M. Pantic, "Diffused heads: Diffusion models beat gans on talking-face generation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 5091–5100.

[13] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3d speaking styles," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 101–10 111.

[14] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "Faceformer: Speech-driven 3d facial animation with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 770–18 780.

[15] A. Richard, M. Zollhöfer, Y. Wen, F. De la Torre, and Y. Sheikh, "Meshtalk: 3d face animation from speech using cross-modality disentanglement," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1173–1182.

[16] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans." *ACM Trans. Graph.*, vol. 36, no. 6, pp. 194–1, 2017.

[17] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *2009 sixth IEEE international conference on advanced video and signal based surveillance*. Ieee, 2009, pp. 296–301.

[18] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao, "Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2020, pp. 601–610.

[19] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 851–866.

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

[21] Y. Hu, C. Luo, and Z. Chen, "Make it move: controllable image-to-video generation with text descriptions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 219–18 228.

[22] S. Stan, K. I. Haque, and Z. Yumak, "Facediffuser: Speech-driven 3d facial animation synthesis using diffusion," in *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games*, 2023, pp. 1–11.

[23] Z. Chen, J. Cao, Z. Chen, Y. Li, and C. Ma, "Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 3, 2025, pp. 2403–2410.

[24] W. Song, X. Wang, S. Zheng, S. Li, A. Hao, and X. Hou, "Talkingstyle: personalized speech-driven 3d facial animation with style preservation," *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[25] H. Wang, Y. Weng, Y. Li, Z. Guo, J. Du, S. Niu, J. Ma, S. He, X. Wu, Q. Hu *et al.*, "Emotivetalk: Expressive talking head generation through audio information decoupling and emotional video diffusion," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 26 212–26 221.

[26] J. Sun, X. Wang, L. Wang, X. Li, Y. Zhang, H. Zhang, and Y. Liu, "Next3d: Generative neural texture rasterization for 3d-aware head avatars," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 20 991–21 002.

[27] Y. Deng, D. Wang, X. Ren, X. Chen, and B. Wang, "Portrait4d: Learning one-shot 4d head avatar synthesis using synthetic data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7119–7130.

[28] T.-C. Wang, A. Mallya, and M.-Y. Liu, "One-shot free-view neural talking-head synthesis for video conferencing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 039–10 049.

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

[30] B. Lin, Y. Yu, J. Ye, R. Lv, Y. Yang, R. Xie, P. Yu, and H. Zhou, "Takin-ada: Emotion controllable audio-driven animation with canonical and landmark loss optimization," *arXiv preprint arXiv:2410.14283*, 2024.

[31] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang, "Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 8652–8661.

[32] Y. Gan, Z. Yang, X. Yue, L. Sun, and Y. Yang, "Efficient emotional adaptation for audio-driven talking-head generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 634–22 645.

[33] J. Cui, H. Li, Y. Zhan, H. Shang, K. Cheng, Y. Ma, S. Mu, H. Zhou, J. Wang, and S. Zhu, "Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 21 086–21 095.

[34] N. Drobyshev, A. B. Casademunt, K. Vougioukas, Z. Landgraf, S. Petridis, and M. Pantic, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8498–8507.

[35] Y. Ma, S. Wang, Z. Hu, C. Fan, T. Lv, Y. Ding, Z. Deng, and X. Yu, "Styletalk: One-shot talking head generation with controllable speaking styles," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 2, 2023, pp. 1896–1904.

[36] B. Liang, Y. Pan, Z. Guo, H. Zhou, Z. Hong, X. Han, J. Han, J. Liu, E. Ding, and J. Wang, "Expressive talking head generation with granular audio-visual control," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3387–3396.

[37] J. Cui, H. Li, Y. Yao, H. Zhu, H. Shang, K. Cheng, H. Zhou, S. Zhu, and J. Wang, "Hallo2: Long-duration and high-resolution audio-driven portrait image animation," *arXiv preprint arXiv:2410.07718*, 2024.

[38] D. Wang, Y. Deng, Z. Yin, H.-Y. Shum, and B. Wang, "Progressive disentangled representation learning for fine-grained controllable talking head synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 979–17 989.

[39] S. Mei, H. Shi, C. Wu, and Z. Chen, "Detailed expression capture and animation model for 3d face recognition," in *2024 5th International Conference on Machine Learning and Computer Application (ICMLCA)*. IEEE, 2024, pp. 306–310.

[40] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "Mead: A large-scale audio-visual dataset for emotional talking-face generation," in *European conference on computer vision*. Springer, 2020, pp. 700–717.

[41] Z. Zhang, L. Li, Y. Ding, and C. Fan, "Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3661–3670.

[42] H. Zhu, W. Wu, W. Zhu, L. Jiang, S. Tang, L. Zhang, Z. Liu, and C. C. Loy, "Celebv-hq: A large-scale video facial attributes dataset," in *European conference on computer vision*. Springer, 2022, pp. 650–667.

[43] W. Zhong, C. Fang, Y. Cai, P. Wei, G. Zhao, L. Lin, and G. Li, "Identity-preserving talking face generation with landmark and appearance priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9729–9738.

[44] C. Liang, Q. Wang, Y. Chen, and M. Tang, "Wav2lip-hr: Synthesising clear high-resolution talking head in the wild," *Computer Animation and Virtual Worlds*, vol. 35, no. 1, p. e2226, 2024.

[45] J. Park and H. Ko, "Real-time continuous phoneme recognition system using class-dependent tied-mixture hmm with hbt structure for speech-driven lip-sync," *IEEE Transactions on Multimedia*, vol. 10, no. 7, pp. 1299–1306, 2008.

[46] Z. Peng, W. Hu, Y. Shi, X. Zhu, X. Zhang, H. Zhao, J. He, H. Liu, and Z. Fan, "Synctalk: The devil is in the synchronization for talking head synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 666–676.

[47] S. J. Park, M. Kim, J. Hong, J. Choi, and Y. M. Ro, "Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 2062–2070.

[48] J. Guan, Z. Zhang, H. Zhou, T. Hu, K. Wang, D. He, H. Feng, J. Liu, E. Ding, Z. Liu *et al.*, "Stylesync: High-fidelity generalized and personalized lip sync in style-based generator," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1505–1515.

[49] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 157–164.

[50] D. Bigioi, S. Basak, M. Stypułkowski, M. Zieba, H. Jordan, R. McDonnell, and P. Corcoran, "Speech driven video editing via an audio-conditioned diffusion model," *Image and Vision Computing*, vol. 142, p. 104911, 2024.

[51] C. Du, Q. Chen, T. He, X. Tan, X. Chen, K. Yu, S. Zhao, and J. Bian, "Dae-talker: High fidelity speech-driven talking face generation with diffusion autoencoder," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 4281–4289.

[52] S. Shen, W. Zhao, Z. Meng, W. Li, Z. Zhu, J. Zhou, and J. Lu, "Difftalk: Crafting diffusion models for generalized audio-driven portraits animation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 1982–1991.

[53] Z. Ma, X. Zhu, G. Qi, C. Qian, Z. Zhang, and Z. Lei, "Diffspeaker: Speech-driven 3d facial animation with diffusion transformer," *arXiv preprint arXiv:2402.05712*, 2024.

[54] Z. Yu, Z. Yin, D. Zhou, D. Wang, F. Wong, and B. Wang, "Talking head generation with probabilistic audio-to-visual diffusion priors," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7645–7655.

[55] J. Chen, Y. Liu, J. Wang, A. Zeng, Y. Li, and Q. Chen, "Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7352–7361.

[56] L. Fan, W. Huang, C. Gan, J. Huang, and B. Gong, "Controllable image-to-video translation: A case study on facial expression generation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 3510–3517.

[57] S. Sinha, S. Biswas, R. Yadav, and B. Bhowmick, "Emotion-controllable generalized talking face generation," *arXiv preprint arXiv:2205.01155*, 2022.

[58] M. Kowalski, S. J. Garbin, V. Estellers, T. Baltrušaitis, M. Johnson, and J. Shotton, "Config: Controllable neural face image generation," in *European Conference on Computer Vision*. Springer, 2020, pp. 299–315.

[59] P. Kumar, K. Deivanai, S. Srivathsav, M. Uthandeeswar, and S. S. Pandi, "A novel approach for face generator based on emotions," in *2024 International Conference on Computational Intelligence for Green and Sustainable Technologies (ICCIGST)*. IEEE, 2024, pp. 1–6.

[60] J. M. Susskind, G. E. Hinton, J. R. Movellan, and A. K. Anderson, "Generating facial expressions with deep belief nets," *Affective computing, emotion modelling, synthesis and recognition*, vol. 2008, no. 5, pp. 421–440, 2008.

[61] E. G. Krumhuber, L. Tamarit, E. B. Roesch, and K. R. Scherer, "Facsgen 2.0 animation software: generating three-dimensional facs-valid facial expressions for emotion research." *Emotion*, vol. 12, no. 2, p. 351, 2012.

[62] H. Chen, H. Zhang, S. Zhang, X. Liu, S. Zhuang, Y. Zhang, P. Wan, D. Zhang, and S. Li, "Cafe-talk: Generating 3d talking face animation with multimodal coarse-and fine-grained control," *arXiv preprint arXiv:2503.14517*, 2025.

[63] X. Luo, S. Takamichi, T. Koriyama, Y. Saito, and H. Saruwatari, "Emotion-controllable speech synthesis using emotion soft labels and fine-grained prosody factors," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 794–799.

[64] S. Zhao, X. Hong, J. Yang, Y. Zhao, and G. Ding, "Toward label-efficient emotion and sentiment analysis," *Proceedings of the IEEE*, vol. 111, no. 10, pp. 1159–1197, 2023.

[65] J. Deng and F. Ren, "A survey of textual emotion recognition and its challenges," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 49–67, 2021.

[66] T. Krishna, A. Rai, S. Bansal, S. Khandelwal, S. Gupta, and D. Goyal, "Emotion recognition using facial and audio features," in *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013, pp. 557–564.

[67] M. Song, J. Bu, C. Chen, and N. Li, "Audio-visual based emotion recognition-a new approach," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 2. IEEE, 2004, pp. II–II.

[68] C. S. Ooi, K. P. Seng, L.-M. Ang, and L. W. Chew, "A new approach of audio emotion recognition," *Expert systems with applications*, vol. 41, no. 13, pp. 5858–5869, 2014.

[69] X. Ji, H. Zhou, K. Wang, W. Wu, C. C. Loy, X. Cao, and F. Xu, "Audio-driven emotional video portraits," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 080–14 089.

[70] A. Metallinou, S. Lee, and S. Narayanan, "Audio-visual emotion recognition using gaussian mixture models for face and voice," in *2008 Tenth IEEE international symposium on multimedia*. IEEE, 2008, pp. 250–257.

[71] B. H. Le, X. Ma, and Z. Deng, "Live speech driven head-and-eye motion generators," *IEEE transactions on visualization and computer graphics*, vol. 18, no. 11, pp. 1902–1914, 2012.

[72] Y. Gao, Y. Zhou, J. Wang, X. Li, X. Ming, and Y. Lu, "High-fidelity and freely controllable talking head video generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 5609–5619.

[73] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, "Gaze estimation from eye appearance: A head pose-free method via eye image synthesis," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3680–3693, 2015.

[74] O. Wiles, A. Koepke, and A. Zisserman, "X2face: A network for controlling face generation using images, audio, and pose codes," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 670–686.

[75] S. Chatzidimitriadis, S. M. Bafti, and K. Sirlantzis, "Non-intrusive head movement control for powered wheelchairs: A vision-based approach," *IEEE Access*, vol. 11, pp. 65 663–65 674, 2023.

[76] L. Chen, G. Cui, C. Liu, Z. Li, Z. Kou, Y. Xu, and C. Xu, "Talking-head generation with rhythmic head motion," in *European conference on computer vision*. Springer, 2020, pp. 35–51.

[77] L. Li, S. Wang, Z. Zhang, Y. Ding, Y. Zheng, X. Yu, and C. Fan, "Write-a-speaker: Text-based emotional and rhythmic talking-head generation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 3, 2021, pp. 1911–1920.

[78] Y. Guo, K. Chen, S. Liang, Y.-J. Liu, H. Bao, and J. Zhang, "Ad-nerf: Audio driven neural radiance fields for talking head synthesis," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5784–5794.

[79] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Fvd: A new metric for video generation," 2019.

[80] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 251–263.

[81] L. Tian, Q. Wang, B. Zhang, and L. Bo, "Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions," in *European Conference on Computer Vision*. Springer, 2024, pp. 244–260.

[82] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3D face model from in-the-wild images," vol. 40, no. 8, 2021. [Online]. Available: https://doi.org/10.1145/3450626.3459936

Supplementary Materials for the EP-Talk Method

1 Implementation Details

1.1 Architecture of the Emotion-Aware Facial Dynamics Module

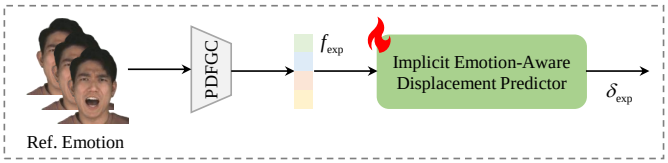

Fig. 1. The architecture of the emotion-aware facial dynamics module.

To accurately extract facial emotional features from videos, we employ the PDFGC method [1] for facial emotion control, as shown in Fig. 1. Given the large number of control parameters extracted from the video, such as identity, pose, and gaze direction, using a joint control approach could lead to feature coupling. Therefore, to avoid feature coupling, we use only emotion-related parameters for facial emotion control when extracting emotional features.

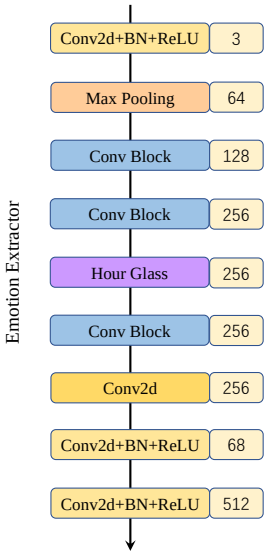

Fig. 2. The network architecture of the emotion feature extractor (PDFGC).

The structure of the PDFGC feature extractor is shown in Fig. 2. It first increases the image dimensions through convolutional layers, then learns emotion-related information via max pooling layers, convolutional layers, and the Hourglass module. Finally, it outputs feature information, such as head pose, facial emotion, gaze direction, and mouth movement, through a linear convolutional layer.

1.2 Architecture of the Motion-Aware Personalized Dynamics Module

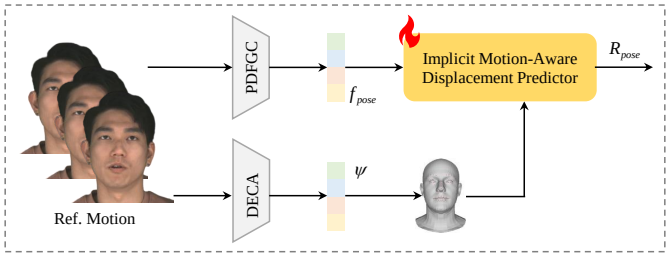

Fig. 3. The architecture of the motion-aware personalized dynamics module.

To estimate personalized head motion coefficients from reference videos, two feature extraction methods, PDFGC and DECA [2], are employed to control personalized head motion characteristics, as illustrated in Fig. 3. The PDFGC method extracts a comprehensive set of features, including head motion, gaze direction, facial expressions, and mouth movements. However, only the head motion parameters are selected for semantic control. In the DECA feature extractor, pose control coefficients are obtained from the input images and subsequently concatenated with the PDFGC features to jointly modulate pose characteristics in the generated video. Given the head motion parameters, the implicit motion-aware displacement predictor estimates the deviation of head keypoint movements, which is then superimposed onto the facial motions during the NTHG stage, ultimately yielding a video with personalized head motion.

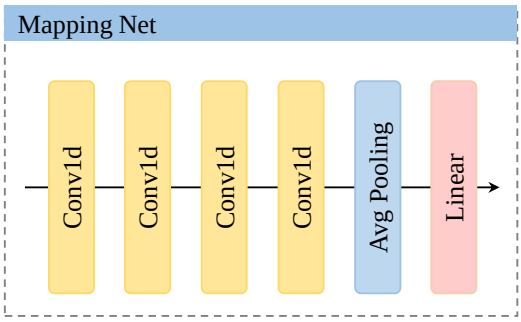

Fig. 4. The architecture of the simple mapping network.

1.3 Architecture of Mapping Network

Considering that the primary task in our experiments is to predict keypoint deviations, specifically predicting a lower-

dimensional parameter space of 21×3, and given the limited experimental data available for training, the designed LADP, EAFM, and MAPD are structured to be as simple and efficient as possible.

The structures of these three mapping networks are similar, with the primary difference lying in the input and output dimensions, as shown in Fig. 4. The extracted features pass through a 1D convolutional layer, an average pooling layer, and a linear layer, ultimately predicting the keypoint deviations. The output dimensions of these three mapping networks differ: the lip region includes keypoints [6, 12, 14, 17, 19, 20]; the facial emotion region includes keypoints [1, 2, 6, 12, 13, 14, 16, 17, 18, 19, 20]; and the head movement region includes keypoints from 0 to 20. Due to the instability in predicting eye gaze direction from audio, we excluded the eye gaze keypoints 11 and 15 during the emotion prediction phase.

1.4 Data Selection and Preprocessing

CelebV-HQ dataset. In the high-quality audio lip synchronization training, we used a pre-trained model, which had been trained on 1,000 videos, as the foundation. Additionally, we selected 100 frontal-view English-speaking videos of different individuals with minimal motion and no occlusion from the CelebV-HQ dataset [3]. These videos were split into training and validation sets in a 4:1 ratio. The model weights were fine-tuned using this carefully selected data, with the primary aim of enhancing the model's performance in audio lip synchronization.

MEAD dataset. The MEAD [4] dataset includes eight different emotion types for several specific individuals: anger, disgust, contempt, fear, happiness, sadness, surprise, and neutrality. Each emotion type is categorized into three levels, ranging from mild to intense, with approximately 30-40 video clips for each level. For our current study, we selected ten commonly used individuals from the MEAD dataset, trained emotion models for each individual, and then conducted experiments in cross-identity reenactment settings.

HDTF dataset. The HDTF dataset [5] contains numerous YouTube video links. From these links, we select high-quality video segments featuring distinct head movements. Given that the videos are relatively long, we trim them into approximately 10-second clips. Subsequently, we divide the clips into training and validation sets in a 4:1 ratio, with each individual's motion video typically lasting between 1 and 3 minutes.

1.5 Inference

During the inference stage, the A2ID, I2F, and LADP modules form the foundational components for natural audio-driven head video generation, enabling high-quality audio synchronization. The EAFD module is responsible for controlling facial emotions, while the MAPD module focuses on personalized head movement generation. These modules can be controlled independently or work in conjunction. The different types of features predicted by these models are aligned in the latent space, avoiding coupling between distinct attributes and thereby enabling the generation of talking videos with stylized feature expressions.

1.6 Several baseline methods for comparison

To accurately evaluate the model's ability to control emotions and head movements, we selected various driving methods, including text labels, speech features, and reference videos. Additionally, we compared our approach with classical audio-driven talking face generation methods.

SadTalker. SadTalker [6] introduces a novel approach that generates realistic talking head videos by learning 3D motion coefficients from audio, explicitly modeling facial expressions and head poses, and mapping them to a 3D-aware face render for high-quality, coherent video synthesis.

Diffused heads. This work [7] introduces an autoregressive diffusion model that generates realistic talking head videos from a single identity image and audio sequence, effectively synthesizing head movements and facial expressions while maintaining the background, achieving SOTA performance on multiple datasets.

EchoMimic. EchoMimic [8] introduces a novel approach that concurrently leverages both audio and facial landmarks to generate dynamic portrait videos, overcoming the limitations of previous methods and achieving superior performance in both quantitative and qualitative evaluations.

AniPortrait. AniPortrait [9] introduces a novel framework that generates high-quality, temporally consistent portrait animations driven by audio and a reference portrait image, achieving superior facial naturalness, pose diversity, and visual quality, with significant potential for flexible and controllable applications in facial motion editing and reenactment.

Hallo2. Hallo2 [10] introduces significant advancements in portrait image animation by enabling long-duration, 4K resolution, audio-driven video synthesis with enhanced temporal coherence, visual consistency, and controllability through the incorporation of textual prompts, achieving SOTA performance on multiple datasets.

EAMM. This paper [11] introduces the Emotion-Aware Motion Model (EAMM), which generates one-shot emotional talking faces by incorporating an emotion source video and leveraging an Audio2Facial-Dynamics module along with an Implicit Emotion Displacement Learner to produce realistic, emotion-driven facial dynamics for arbitrary subjects.

EAT. The EAT method [12] introduces parameter-efficient adaptations to transform emotion-agnostic talking-head models into emotion-controllable ones, achieving SOTA performance in emotion control and generalization, even with limited emotional training data.

IP_LAP. This paper [13] proposes a two-stage framework for generating realistic, lip-synced, and identity-preserving talking face videos from audio, utilizing a transformer-based landmark generator and a video rendering model that incorporates prior appearance and auditory features to improve synchronization and realism.

Hallo3. This paper [14] presents a pretrained transformer-based generative video model for portrait animation. The model is designed to handle non-frontal viewpoints, dynamic objects, and immersive backgrounds, while preserving facial identity consistency and generating realistic audio-driven videos.

2 Comparison of Experimental Results

2.1 Comparison of methods for controlling facial emotions

TABLE 1
Comparison of experimental results for audio-driven identity reconstruction on the MEAD dataset.

Method	FID ↓	FVD ↓	Sync-D ↓	E-FID ↓
SadTalker [6]	108.1	5.828	11.68	0.221
EchoMimic [8]	77.72	5.210	11.94	0.117
AniPortrait [9]	107.6	5.695	12.83	0.179
Hallo2 [10]	68.79	4.862	10.76	0.173
EAMM [11]	122.8	7.364	12.18	0.293
EAT [12]	64.82	4.351	10.82	0.352
Hallo3 [14]	75.76	5.948	10.40	0.180
EP-Talk	51.40	4.029	10.67	0.075

To evaluate the model's reconstruction performance, we compare our approach with the latest methods, as shown in Fig. 10. Among these, SadTalker, Echomimic, AniPortrait, and Hallo2 take a reference portrait and driving audio as input. In contrast, EAMM, EAT, Hallo3, and our method require additional emotional video input alongside the reference portrait and driving audio. Methods such as EAMM and Hallo3 have limitations in preserving the identity of specific individuals. Furthermore, SadTalker, Echomimic, AniPortrait, and EAT exhibit noticeable blurring in the mouth region. More critically, the facial emotion features reconstructed by these methods are inaccurate, with significant discrepancies when compared to the reference emotional video. Extensive experimental results demonstrate that our proposed method shows clear advantages in image generation quality and facial emotion control.

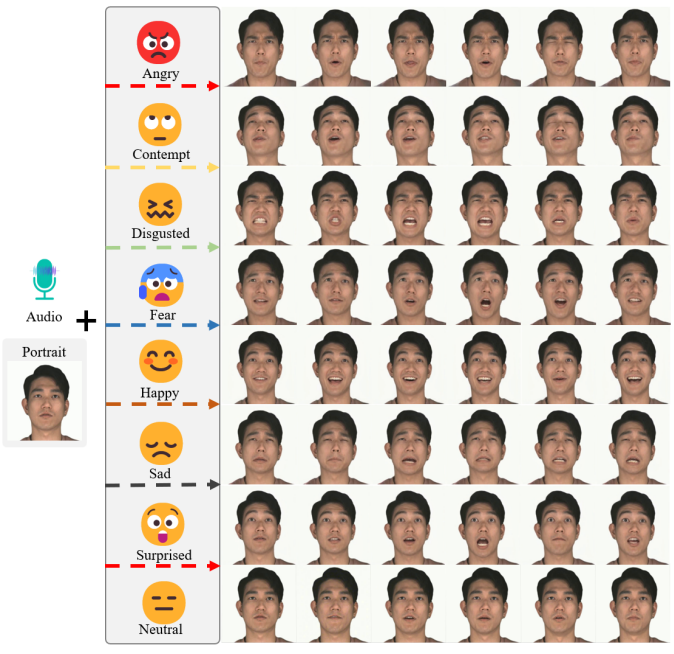

Fig. 5. Experimental results for eight different facial expressions of Individual 1 on the MEAD dataset.

The quantitative experimental results for reconstruction on the MEAD dataset are presented in Tab. 1. Our method achieves the best experimental results in terms of image quality (FID, FVD) and facial emotion control (E-FID), and delivers performance comparable to Hallo3 in lip synchronization.

To validate the effectiveness of the proposed method in controlling facial emotions, we performed reconstructions using eight different emotions across multiple individuals, with the experimental results shown in Fig. 5 and Fig. 6. Given the input audio, reference portrait, and emotional video, our method accurately reconstructs various facial emotions while maintaining better identity consistency.

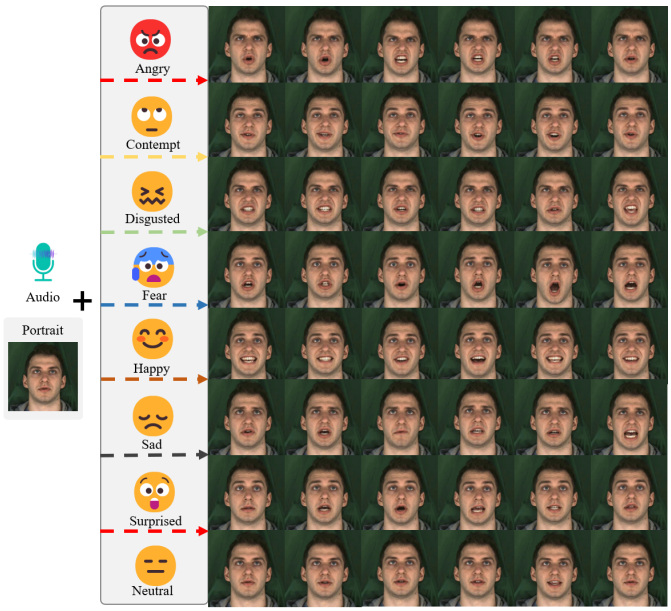

Fig. 6. Experimental results for eight different facial expressions of Individual 2 on the MEAD dataset.

2.2 Comparison of methods for personalized head movement control

TABLE 2
Comparison of experimental results for audio-driven identity reconstruction on the HDTF dataset.

Method	FID ↓	FVD ↓	Sync-D ↓	E-FID ↓
SadTalker [6]	49.07	3.173	13.86	0.055
EchoMimic [8]	29.31	2.687	12.93	0.058
AniPortrait [9]	57.31	4.724	13.40	0.052
Hallo2 [10]	25.49	2.227	11.45	0.034
EAMM [11]	43.20	6.191	13.26	0.106
EAT [12]	46.59	5.859	14.16	0.142
Hallo3 [14]	23.91	3.149	10.39	0.194
EP-Talk	20.97	1.997	10.85	0.030

To evaluate the effectiveness of head motion reconstruction, we conducted experiments on the HDTF dataset and compared our method with SOTA approaches. The experimental results are shown in Fig. 11, where we compare the

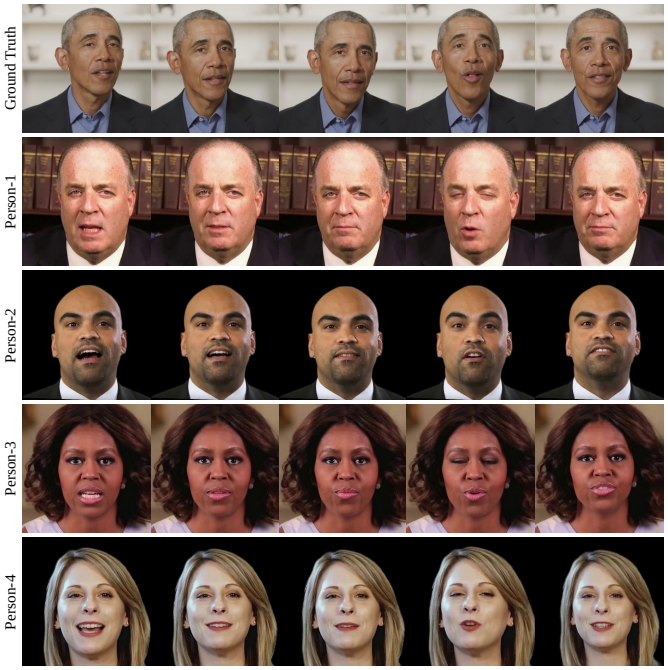

Fig. 7. Experimental results of cross-identity reenactment on the HDTF dataset.

motion features of two individuals and provide the corresponding ground truth results in the last row. The images generated by methods such as AniPortrait, Hallo2, EAMM, EAT, and Hallo3 exhibit noticeable blurring or artifacts. The images produced by SadTalker lack facial detail information. While Echomimic, EAT, and Hallo3 generate noticeable motion, their motion styles do not align with those of the reference video. We hypothesize that these methods focus more on the consistency between head movements and audio rhythm, neglecting the personalized head movements present in the reference video. The results demonstrate that the proposed method, EP-talk, reconstructs head motions more accurately and effectively learns personalized head movements for specific individuals.

The quantitative reconstruction results on the MEAD dataset are reported in Tab. 2. Our method achieves the best performance in terms of image quality and facial emotion control, and attains audio–lip synchronization comparable to that of Hallo3. Since Hallo2 and Hallo3 place greater emphasis on learning the relationship between audio prosody and facial expressions, they exhibit stronger performance in mouth-shape synchronization and facial emotion rendering.

We perform cross-identity reenactment on the HDTF dataset, using head movements from different individuals as references to generate personalized head movement styles tailored to specific identities, as shown in Fig. 7. Experimental results indicate that using personalized head motion styles for different individuals leads to noticeable differences in the cross-identity driven generation outcomes.

2.3 Ablation results for facial emotion and head movement

We conduct ablation experiments on the HDTF and MEAD datasets for reconstruction, with quantitative results presented in Tab. 3 and qualitative results shown in Figs. 8 and 9. Fig. 8 illustrates the ablation results for facial emotion, while Fig. 9 shows the ablation results for head movement. "w/o Natural" refers to the absence of the I2F and LADP modules, "w/o Emotion-Aware" indicates the lack of the EAFD module, and "w/o Motion-Aware" denotes the omission of the MAPD module.

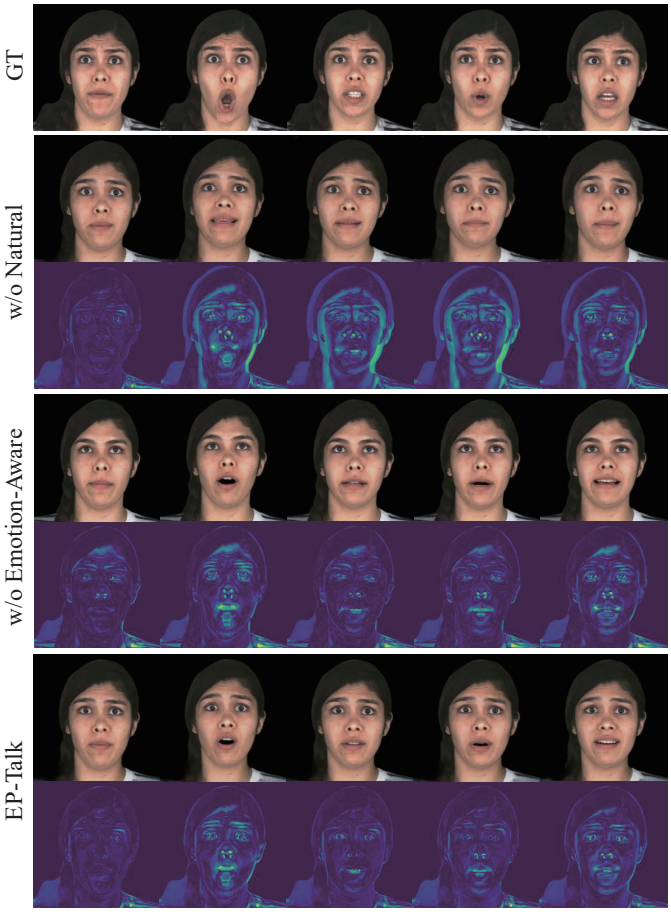

Fig. 8. Ablation study results for the NTHG module and the EAFD module. The heat map beneath each image highlights discrepancies between the generated image and the reference expression image, where yellow denotes significant errors and blue indicates minimal differences.

TABLE 3
Ablation study of EP-Talk. "*" denotes results on the HDTF dataset, while unmarked results are from MEAD.

Method	FID ↓	FVD ↓	Sync-D ↓
w/o Natural	93.34	11.84	15.04
w/o Emotion-Aware	96.59	5.822	10.75
EP-Talk	51.40	4.029	10.67
w/o Motion-Aware*	23.33	2.253	9.026
EP-Talk*	20.45	2.085	10.85

As shown in Fig. 8, the absence of the natural talking module results in overall blurring in the generated images, while the lack of the EAFD module leads to noticeable deviations in facial features. In Fig. 9, the absence of the MAPD module causes the head region to remain almost unchanged, and the generated video lacks stylized head

motion characteristics. The ablation study results in Tab. 3 are consistent with the quantitative experimental results, where the absence of the emotion and motion modules significantly reduces image quality.

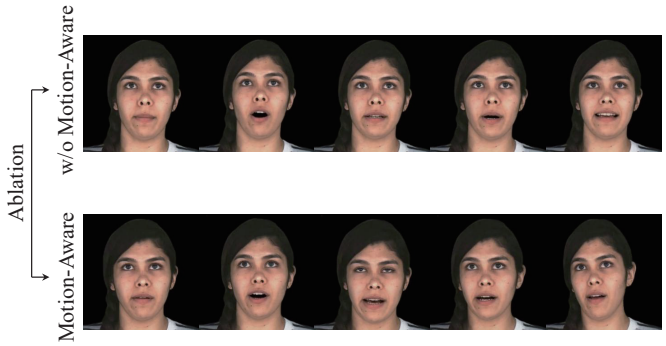

Fig. 9. Evaluation of the MAPD module in the ablation study.

REFERENCES

[1] D. Wang, Y. Deng, Z. Yin, H.-Y. Shum, and B. Wang, "Progressive disentangled representation learning for fine-grained controllable talking head synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 979–17 989.

[2] S. Mei, H. Shi, C. Wu, and Z. Chen, "Detailed expression capture and animation model for 3d face recognition," in *2024 5th International Conference on Machine Learning and Computer Application (ICMLCA)*. IEEE, 2024, pp. 306–310.

[3] H. Zhu, W. Wu, W. Zhu, L. Jiang, S. Tang, L. Zhang, Z. Liu, and C. C. Loy, "Celebv-hq: A large-scale video facial attributes dataset," in *European conference on computer vision*. Springer, 2022, pp. 650–667.

[4] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "Mead: A large-scale audio-visual dataset for emotional talking-face generation," in *European conference on computer vision*. Springer, 2020, pp. 700–717.

[5] Z. Zhang, L. Li, Y. Ding, and C. Fan, "Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3661–3670.

[6] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang, "Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 8652–8661.

[7] M. Stypułkowski, K. Vougioukas, S. He, M. Zieba, S. Petridis, and M. Pantic, "Diffused heads: Diffusion models beat gans on talking-face generation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 5091–5100.

[8] Z. Chen, J. Cao, Z. Chen, Y. Li, and C. Ma, "Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 3, 2025, pp. 2403–2410.

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

[10] J. Cui, H. Li, Y. Yao, H. Zhu, H. Shang, K. Cheng, H. Zhou, S. Zhu, and J. Wang, "Hallo2: Long-duration and high-resolution audio-driven portrait image animation," *arXiv preprint arXiv:2410.07718*, 2024.

[11] X. Ji, H. Zhou, K. Wang, Q. Wu, W. Wu, F. Xu, and X. Cao, "Eamm: One-shot emotional talking face via audio-based emotion-aware motion model," in *ACM SIGGRAPH 2022 conference proceedings*, 2022, pp. 1–10.

[12] Y. Gan, Z. Yang, X. Yue, L. Sun, and Y. Yang, "Efficient emotional adaptation for audio-driven talking-head generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 634–22 645.

[13] W. Zhong, C. Fang, Y. Cai, P. Wei, G. Zhao, L. Lin, and G. Li, "Identity-preserving talking face generation with landmark and appearance priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9729–9738.

[14] J. Cui, H. Li, Y. Zhan, H. Shang, K. Cheng, Y. Ma, S. Mu, H. Zhou, J. Wang, and S. Zhu, "Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 21 086–21 095.

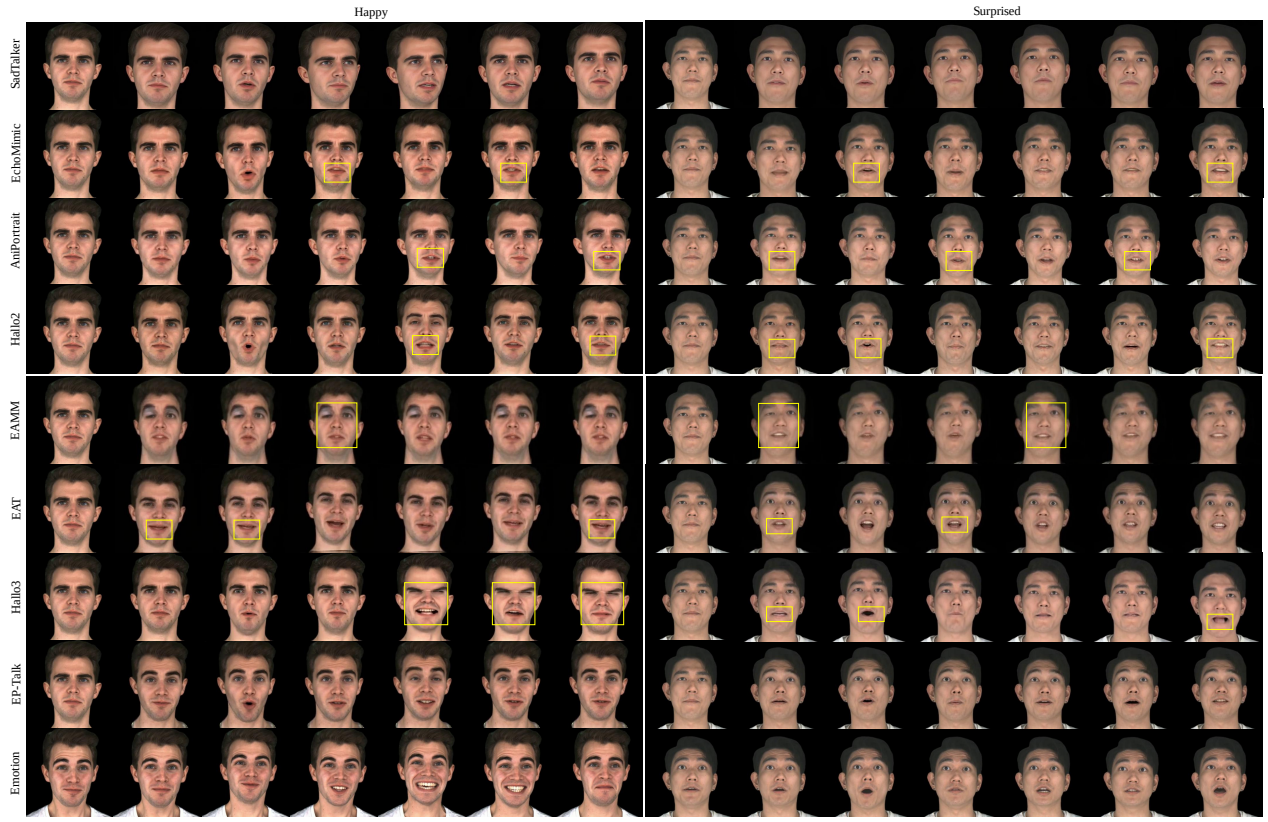

Fig. 10. Comparison of experimental results for audio-driven reconstruction on the MEAD dataset. Identity features from two individuals are used for reconstruction. Each row in the figure represents the reconstruction results, while the last row shows the reference facial expressions.

Fig. 11. Comparison of experimental results for audio-driven reconstruction on the HDTF dataset. Identity features from two individuals are used for reconstruction. Each row in the figure represents the reconstruction results, while the last row shows the reference head movements.