Takin-ADA: Emotion Controllable Real-Time Audio-Driven Animation with Canonical and Landmark Loss Optimization

Anonymous cvm submission

Paper ID ****

Abstract

We present *Takin-ADA*, which enables real-time audio-driven animation of individual portraits utilizing 3D implicit keypoints, while also allowing for precise control over facial expressions for the first time. Takin-ADA tackles critical issues faced by existing audio-driven facial animation methods, notably expression leakage, subtle expression transfer and audio-driven precision through a two-stage approach. In the first stage, we ingeniously incorporate a canonical loss and a landmark-guided loss to enhance the transfer of subtle expressions while simultaneously mitigating expression leakage. These advancements significantly elevate the quality and realism of the generated facial animations. The second stage employs a diffusion model framework leveraging HuBERT features, which substantially improves lip-sync accuracy, ensuring a more natural and synchronized audio-visual experience. Through this two-stage approach, Takin-ADA not only generates precise lip movements but also allows flexible control over expression and head motion parameters, resulting in more natural and expressive facial animations. Takin-ADA is capable of generating high-resolution facial animations in real-time, outperforming existing commercial solutions. Extensive experiments demonstrate that our model significantly surpasses previous methods in various aspects, including video quality, facial dynamics realism, and naturalness of head movements.

Keywords: Audio-Driven Portraits Animation, Two-Stage, 3D Implicit Keypoints, Canonical Loss, Diffusion Model, expression control

In recent years, portrait animation has emerged as a pivotal area of research in computer vision, driven by its wide-ranging applications in digital human animation, film dubbing, and interactive media[34, 23, 59]. The ability to generate realistic, expressive, and controllable facial animations from a single image has become increasingly important in creating lifelike digital avatars for various applications, including virtual hosts, online education, and digital human interactions[28, 49, 29].

Existing approaches to portrait animation can be broadly categorized into two paradigms: audio-driven[40, 34, 59, 57, 60, 61] and video-driven animation[45, 44, 17]. While these methods have shown promise, they face significant challenges in achieving precise control over facial expressions, maintaining identity consistency, and generating natural head movements. Audio-driven methods often struggle to capture the full spectrum of non-verbal cues, resulting in animations that lack expressiveness[62, 43, 51]. Video-driven techniques, while potentially capturing a wider range of facial dynamics, often suffer from expression leakage, where the source video's expressions unduly influence the animated output[45, 40].

The primary challenge in this field lies in developing a unified framework that can simultaneously achieve individual facial control, handle both audio-driven and video-driven talking face generation efficiently, and operate in real-time. Existing models often rely on explicit structural representations such as blendshapes[6, 13, 33] or 3D Morphable Models (3DMM)[9, 14, 30], which offer constrained approximations of facial dynamics and fail to capture the full breadth of human expressiveness.

To address these limitations, we present Takin-ADA (Audio-Driven Animation), an innovative two-stage framework for real-time audio-driven animation of single-image portraits with controllable expressions using 3D implicit keypoints[44]. Our approach tackles the critical issues of expression leakage, subtle expression transfer, and audio-driven precision through a carefully designed two-stage process.

In the first stage, we introduce a novel 3D Implicit Keypoints Framework that effectively disentangles motion and appearance. This stage employs a standard face mean absolute error (MAE) loss to mitigate expression leakage and a landmark-based wing loss to enhance the transfer of subtle expressions. These innovations significantly improve the quality and realism of generated facial animations while maintaining identity consistency.

The second stage employs an advanced, audio-

*These authors contributed equally to this work.
†Corresponding author.

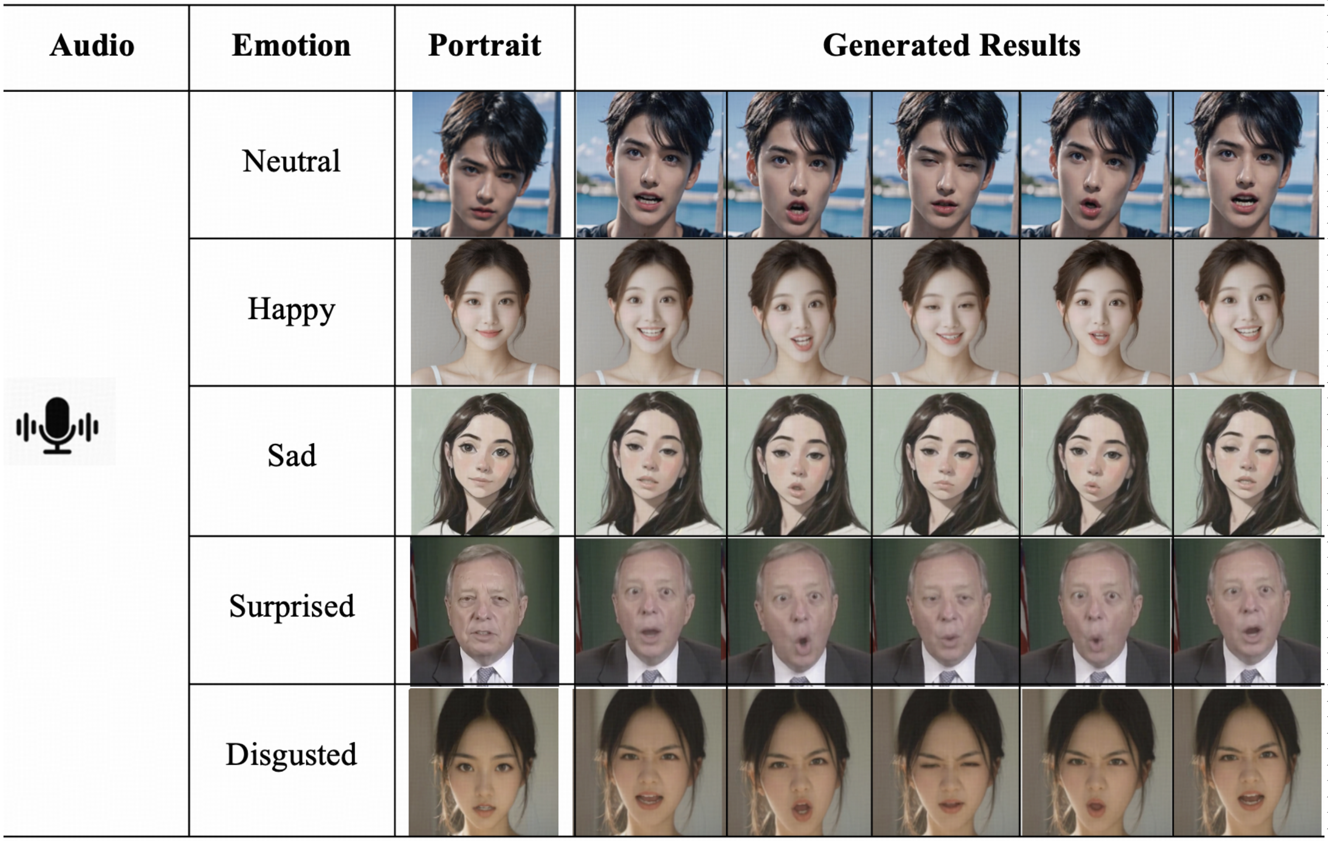

Figure 1. We introduce Takin-ADA, a framework that transforms input audio and a single static portrait into animated talking videos with naturally flowing movements. Each column of generated results utilizes identical control signals with different and expressions but incorporates some random variations, demonstrating the diversity of our generated outcomes.

conditioned diffusion model utilizing HuBERT features. This model not only dramatically improves lip-sync accuracy but also allows for flexible control over expression and head motion parameters. By incorporating a weighted sum technique, our approach achieves unprecedented accuracy in lip synchronization, establishing a new benchmark for realistic speech-driven animations.

A key feature of Takin-ADA is its ability to generate high-resolution facial animations in real-time. Using native pytorch inference on an RTX 4090 GPU, our method achieves the generation of 512×512 resolution videos at up to 42 FPS, from audio input to final portrait output. This breakthrough in efficiency opens new possibilities for real-time digital human interaction and virtual reality applications.

Through extensive experiments and evaluations, we demonstrate that Takin-ADA significantly surpasses previous methods in various aspects, including video quality, facial dynamics realism, and naturalness of head movements. Our comprehensive performance enhancements not only advance the field of digital human technology but also pave the way for creating more natural and expressive AI-driven virtual characters.

In summary, Takin-ADA represents a significant step forward in single-image portrait animation, offering both technological advancements and practical applicability in real-world scenarios. By addressing the critical aspects of audio-driven avatar synthesis, our work provides a solid foundation for future research in this field and has the potential to profoundly impact various domains, including human-computer interaction, education, and entertainment.

1. Related Work

1.1. 3D Implicit Keypoints and Disentangled Face Representation

The representation of facial images has been extensively studied by previous works. Traditional methods employ sparse keypoints[36, 52] or 3D face models[35, 15, 54] to explicitly characterize facial dynamics and other properties. However, these approaches often encounter issues such as inaccurate reconstructions and limited expressive capabilities. Recent advancements have focused on learning disentangled representations within a latent space. A common strategy involves separating faces into identity and non-identity components, which are then recombined across different frames in either 2D or 3D contexts[2, 60, 27, 50, 44, 10]. The primary challenge for these methods lies in effectively disentangling various factors while maintaining expressive representations of all static and dynamic facial attributes. Non-diffusion-based models have employed implicit keypoints as intermediate motion representations, warping the source portrait with the driving image through optical flow. Methods such as FOMM[36] approximate local motion using first-order Taylor expansion near each keypoint and local affine transformations, whilst MRAA utilizes PCA-based motion estimation to represent articulated motion[37]. Face vid2vid[44] extended the FOMM framework by introducing 3D implicit keypoints representation, achieving free-view portrait animation. Despite these advancements, Face vid2vid has limitations in the transfer of subtle expressions.

To address these challenges, several methods have been proposed to improve the warping mechanism and representation of complex motions. IWA enhanced the warping mechanism using cross-modal attention, which can be extended to multiple source images[31]. TPSM employed nonlinear thin-plate spline transformations to estimate optical flow more flexibly and handle large-scale motions more effectively[58]. DaGAN leveraged dense depth maps to estimate implicit keypoints capturing critical driving movements[24] . MCNet introduced an identity representation conditioned memory compensation network to mitigate ambiguous generation caused by complex driving motions[22]. Our work builds upon Face vid2vid[44] by developing a series of significant enhancements to improve expression generalization and expressiveness. Our innovative use of 3D implicit keypoints forms the foundation of the Takin-ADA framework, leading to more accurate and expressive facial animations.

1.2. Audio-Driven Talking Face Generation

Audio-driven talking face generation has been a long-standing challenge in computer vision and graphics. Early efforts primarily focused on synthesizing lip movements from audio signals, leaving other facial attributes unchanged[39, 4, 34]. Recent advancements have expanded the scope to include a broader range of facial expressions and head movements derived from audio inputs. For instance, some methods separate generation targets into categories such as lip-only 3DMM coefficients, eye blinks, and head poses, while others decompose lip and non-lip features on top of expression latents[56]. These approaches typically regress lip-related representations directly from audio features and model other attributes probabilistically[51]. In contrast, our Takin-ADA framework generates comprehensive facial dynamics and head poses from audio along with other control signals, offering a more holistic and integrated approach to audio-driven animation.

1.3. Diffusion Models in Facial Animation

Diffusion models[21] have shown remarkable performance across various generative tasks, including their application as rendering modules in facial animation[12, 18]. While these models often produce high-quality images, they require extensive parameters and substantial training data. To enhance generation efficiency, recent approaches have employed diffusion models for generating motion representations[1, 19]. Diffusion models excel at addressing the one-to-many mapping challenge crucial for speech-driven generation tasks, where the same audio clip can lead to different actions across individuals or even within the same person. The training and inference phases of diffusion models, which systematically introduce and then remove noise, allow for the incorporation of controlled variability during generation. In Takin-ADA, we leverage a state-of-the-art audio-conditioned diffusion model that integrates facial expression and head motion parameters, enabling diverse and controllable facial animations while maintaining high accuracy in lip synchronization.

1.4. Real-Time High-Resolution Video Generation

While recent advancements in image and video diffusion techniques have significantly improved talking face generation[41, 26], their substantial computational demands have limited their practicality for interactive, real-time systems. Our work addresses this critical gap by developing a method that delivers high-quality video output while supporting real-time generation. Takin-ADA achieves the generation of 512×512 resolution videos at up to 42 FPS, from audio input to final portrait output, representing a significant advancement in the field of real-time, high-resolution facial animation.

By addressing these key areas, our Takin-ADA framework represents a comprehensive approach to audio-driven avatar synthesis, combining advanced 3D implicit keypoint representation, sophisticated audio-conditioned diffusion modeling, and efficient real-time generation capabilities.

2. METHODOLOGY

Figure 2 illustrates the structure of Takin-ADA, which takes a single face image of any identity and an arbitrary speech audio clip as input to generate a realistic synthesized video of the input face speaking the given audio. This section elaborates on our method in detail. We start with a brief overview of the Takin-ADA framework. Next, we describe our meticulously designed approach for constructing the latent space of the face. Finally, we introduce our comprehensive system for generating dynamic facial movements.

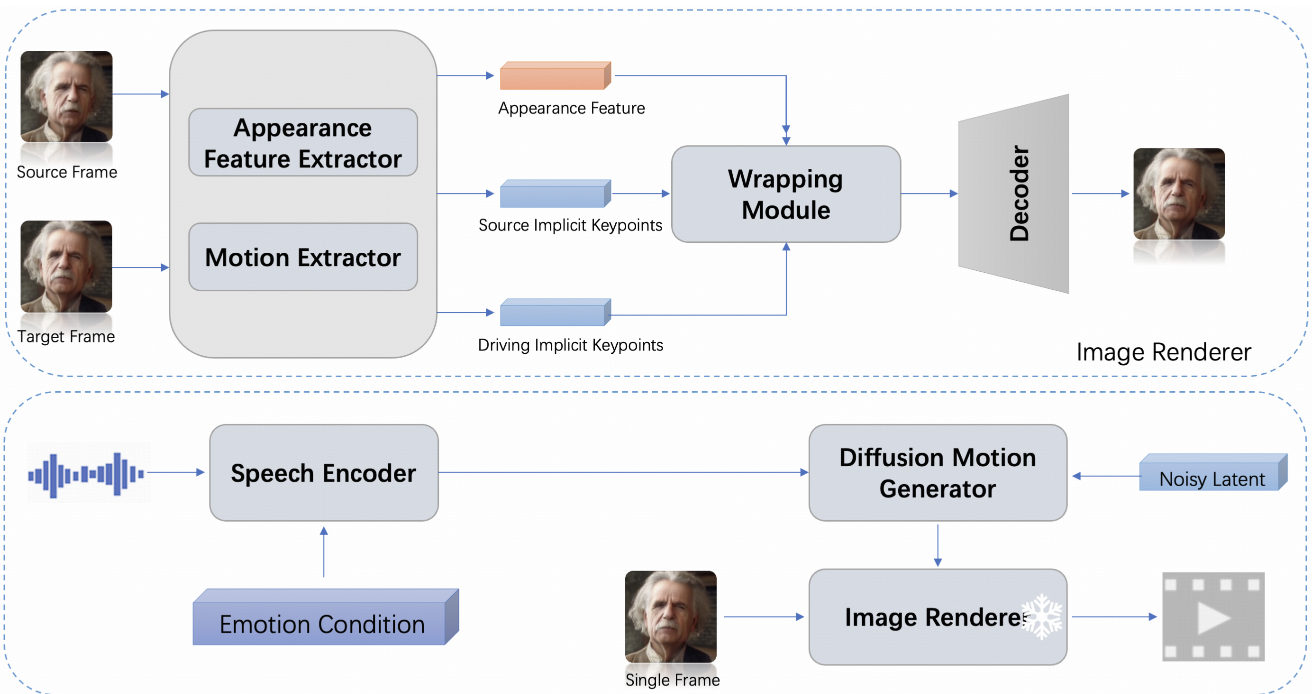

Figure 2. Illustration of our proposed Takin-ADA. The framework comprises two primary components: (1) a representation learning module for extracting expressive and disentangled facial latent representations, and (2) a sequence generation module that synthesizes motion sequences based on audio input. The first component focuses on learning robust motion representations through the utilization of canonical keypoint loss and landmark guidance. Subsequently, these learned motion representations serve as input for the second component, enabling further audio-drive facial image generation and manipulation

2.1. Takin-ADA Framework

Rather than directly generating video frames, we produce holistic facial dynamics and head motion in latent space, conditioned on audio and other signals. These motion latent codes are then used by a face decoder to create video frames, incorporating appearance and identity features extracted from the input image by a face encoder. As illustrated in Figure 2, Takin-ADA encompasses two key components:

- a facial motion representation system capable of capturing universal facial dynamics.
- a face latent generation using user-controlled driving signal to produce the synthesised talking face video.

2.2. Expressive and Disentangled Face Latent Space Construction

In the first-stage, to build a face latent space with high degrees of expressiveness and disentanglement, our approach utilizes a corpus of unlabeled talking face videos in a self-supervised image animation framework which employs a source image I_s and a target image I_t from the same video clip, where I_s provides identity information, I_t delivers motion details. The primary aim of our system is to reconstruct I_t. We choose face vid2vid[44] as our base model to get facial motion latent. Compared to extant facial motion representation methodologies, including blendshapes, landmark coefficients, 2D latent and 3D Morphable Models (3DMM), the trainable latent 3D keypoints demonstrate substantial superiority in capturing nuanced emotional states and subtle facial deformations, thus providing a more sensitive and precise framework for facial animation.These 3D keypoints can be divided into two categories: one that captures facial expressions and another represents an individual's geometric signature which we called canonical volume. The 3D appearance feature volume surpassing 2D feature maps at detailing appearance. Additionally, explicit 3D feature warping proves highly effective in modeling head and facial movements in a 3D space.The source 3D keypoints x_s and the driving 3D keypoints x_d are transformed as follows:

$$\begin{cases} x_s = x_{c,s} R_s + \delta_s + t_s, \\ x_d = x_{c,s} R_d + \delta_d + t_d, \end{cases}$$

where x_s and x_d are the source and driving 3D implicit keypoints, respectively, and $x_{c,s}$represents the canonical keypoints of the source image. The source and driving poses are R_s and R_d, the expression deformations are δ_s and δ_d, and the translations are t_s and t_d .

Significantly, we introduce a suite of pivotal advancements in latent 3D keypoint technology, encompassing canonical volume representation and landmark-guided optimization.

Canonical Keypoints. Although the canonical volume in Takin-ADA was designed to exclude facial expression details, we discovered that the generated expression is heavily influenced by the source image, indicating that information leakage affects image synthesis.Thus, a more neutral canonical volume enhances both tractability and effectiveness in expression translation tasks. To address this problem, we propose matching canonical keypoints from different images of the same person during training, using the following loss function:

$$\mathcal{L}_{canonical} = \frac{1}{N}\sum_{1}^{N}(\mathcal{L}_{Huber}(x_{cs_i}, x_{cs_j})) \tag{1}$$

where x_{cs_i} and $x_{cs_j}$are the canonical keypoints derived from distinct images depicting the same individual. The loss serves to maintain the stability and expression-invariance of the canonical volume, which is paramount for the accurate translation of intense facial expressions.

Landmark Guidance. The original face vid2vid approach [44] appears to have limitations in vividly animating subtle facial expressions. We posit that these shortcomings primarily stem from the inherent challenges of learning nuanced facial expressions through unsupervised methods.Drawing inspiration from [17], we introduce 2D landmarks that capture micro-expressions, using them to guide and optimize the learning of implicit points. The landmark-guided loss $\mathcal{L}_{land}$ is formulated as follows:

$$\mathcal{L}_{landmark} = \frac{1}{2N}\sum_{1}^{N}(\mathcal{L}_{Huber}(l_i, x_{s,i,:2}) + \mathcal{L}_{Huber}(l_i, x_{d,i,:2})) \tag{2}$$

where N is the number of selected landmarks, $x_{s,i,:2}$ and $x_{d,i,:2}$ denote the first two spatial dimensions of the implicit keypoints for source and driving image respectively,Huber loss is adopted following [5].

2.3. Emotional Holistic Facial Motion Generation

After completing the training of the motion encoder and image renderer, we freeze these models and move on to the second phase, which is driven by audio to produce motion conditioned on the audio input. Crucially, we consider holistic facial dynamics generation, where our learned latent codes represent all facial movements such as lip motion, expression, and eye gaze and blinking. Specifically, we employ a combination of diffusion and condition: the diffusion learns a more accurate distribution of motion data, while the emotion condition primarily facilitates attribute manipulation.The trained generative model generates videos that synchronize with the speech signal or other control signals to animate a source image I_s.

Diffusion formulation. Specifically, we employ a multi-layer Conformer[16] for our sequence generation task. Diffusion models utilize two Markov chains: the forward chain progressively adds Gaussian noise to the target data, while the reverse chain iteratively restores the raw signal from this noise. During training, we integrate the diffusion process, where the noising phase gradually transforms clean Motion Latents M into Gaussian noise M^T over a series of denoising steps. Conversely, the denoising phase systematically removes noise from the Gaussian noise[21], ultimately yielding clean Motion Latents. This iterative process better captures the distribution of motion, enhancing the diversity of the generated results.

$$L_{diff} = \mathbb{E}_{t,M,\varepsilon}[\|\varepsilon - \hat{\varepsilon}_t(M_t, t, C)\|^2] \tag{3}$$

Weighted Sum. To enhance the robustness of the audio encoder, we employ a novel approach that retrieves the audio latent code through a weighted summation of all layers within the self-supervised models. This methodology diverges from the conventional Mel-based feature representation, thereby conferring enhanced language flexibility to the system. This approach ensures that the DDIM [38] generates deterministic and consistent outcomes, thus bolstering the reliability and reproducibility of the results.

Emotion Condition. To achieve better performance, we also incorporate emotional condition into the Conformer to enhance facial expressions. Motivated by the observation that variations in facial expressions in a video sequence are generally less frequent than other types of motion changes, we define a window of size K around I_d and average the K extracted expression features to obtain a refined expression feature. This clean expression feature is then combined with the extracted mouth and pose features as input to the generator model. During the inference phase, we can generate videos exhibiting diverse emotional states by assigning different affective vectors to the same audio input. This approach enables the production of emotionally varied outputs from a single audio source. Furthermore, we can leverage the emotional content inherent in the audio to generate videos with enhanced emotional controllability. This method allows for a more nuanced and precise manipulation of the emotional characteristics in the synthesized video output.

3. Experiments

3.1. Experiment Settings

As shown in Table 1, we first give a brief summary of the key features of the existing methods.Next, we give an overview of the implementation details, dataset, benchmarks, and baselines used in the experiments. Then, we

Method	Head Motion	Emotion	HD	Real Time
MakeItTalk[62]	✗	✗	✗	✗
SadTalker[56]	✓	✗	✗	✗
IP_ LAP[54]	✗	✗	✗	✗
AniTalker[28]	✓	✗	✗	✓
EDTalk[40]	✓	✓	✗	✓
EchoMimic[7]	✗	✗	✓	✗
Takin-ADA	✓	✓	✓	✓

Table 1. Summary of Different Portrait Animation Methods

present the experimental results on video-driven methods both self-reenactment and cross-reenactment, and audio-driven methods followed by an ablation study to validate the effectiveness of the proposed calonical keypoint and landmark gudiance.

Implementation Details. The first training phase was conducted using a cluster of eight NVIDIA A800 GPUs over a 8-day period, with models initialized from scratch. Input images were preprocessed through alignment and cropping to a standardized 256×256 pixel resolution. We implemented a batch size of 104 to optimize computational efficiency, while the output resolution was set at 512×512 pixels. We follow *Face Vid2Vid* [44] to use implicit keypoints equivariance loss $\mathcal{L}_E$, keypoint prior loss $\mathcal{L}_L$, head pose loss $\mathcal{L}_H$, and deformation prior loss $\mathcal{L}_\Delta$. To further improve the expression disentanglement, we apply Canonical Keypoints losses and Landmark Guidance losses, denoted as $\mathcal{L}_{\text{canonical}}$ and $\mathcal{L}_{\text{landmark}}$. To further improve the texture quality, we also apply perceptual and GAN losses on the global region of the input image fine-tuned from *LivePortrait* model. In the second phase, the speech encoder and the Motion Generator utilize a four-layer and an eight-layer conformer architecture, respectively, inspired by [11]. This architecture integrates the conformer structure and relative positional encoding [8, 16]. A pre-trained HuBERT-large model [25] serves as the audio feature encoder, incorporating a downsampling layer to adjust the audio sampling rate from 50 Hz to 25 Hz to synchronize with the video frame rate. The training of the audio generation process spans 125 frames (5 seconds). Detailed implementation specifics and model structures are further elaborated in the supplementary materials.

Dataset. Our study employs three distinct datasets: VoxCeleb[32], HDTF[57], and MEAD[42]. To ensure consistency in data processing, we retrieved the original video files from these sources and implemented a standardized processing methodology across all datasets. Furthermore, we augmented our research with a substantial collection of 4K-resolution portrait videos, comprising approximately 200 hours of talking head footage. In preprocessing this additional data, we segmented extended video sequences into clips not exceeding 30 seconds in duration. To maintain data integrity and focus, we utilized face tracking and recognition technologies to ensure that each clip contains footage of only a single individual. This approach enhances the dataset's suitability for our research objectives and facilitates more accurate analysis.

Benchmarks. To quantitatively measure the visual quality, we figure up the Peak Signal-to-Noise Ratio (PSNR), Structure SIMilarity (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) for the generated videos[47, 55]. Following Wav2Lip[34], Lip-sync Distance (LSE-D) is applied to measure the audiovisual synchronization. For assessing reenactment quality, we employ various metrics including the Frechet Inception Distance (FID) to measure the distributional discrepancy between synthetic and real images[20]. Cosine similarity (CSIM) from a face recognition network quantifies the identity preservation in generated images[3] and Structural Similarity Index (SSIM)[46] . Regarding subjective metrics, we employ the Mean Opinion Score (MOS) as our metric, with 35 participants rating our method based on Lip-sync(LS), Naturalness(N), Resolution(R), and Expression Transfer(ET) .

3.2. Summary of the portrait animation methods

Table 1 summarizes the key features of existing methods in terms of high-quality output (HD), real-time performance, and fine-grained control over different aspects, including head motion and emotion. While other approaches excel in some areas, our method uniquely possesses all these desirable characteristics. This comprehensive capability is made possible by our sophisticated universal motion representation, which enables us to balance quality, efficiency, and control effectively. Our approach thus represents a significant advancement in speech-driven facial animation technology, offering a solution that doesn't compromise on any front.

3.3. Video-driven methods

Quantitative Results. We benchmarked our approach against several leading face reenactment methods, all employing variations of self-supervised learning. The results are presented in Table 1. Due to the inherent challenges and the absence of frame-by-frame ground truth in

Method	Self-Reenactment				Cross-Reenactment		
	FID↓	CSIM↑	LPIPS↓	MOS-ET↑	CSIM↑	LPIPS↓	MOS-ET↑
FOMM[36]	32.935	0.825	0.021	2.769	0.174	0.218	1.934
StyleHEAT[50]	33.136	0.522	0.095	2.675	0.244	0.213	1.768
LIA[45]	28.008	0.834	0.021	3.187	0.149	0.216	2.937
FADM[53]	28.981	0.832	0.024	2.763	0.106	0.199	2.268
Face Vid2Vid[44]	28.444	0.831	0.023	3.451	0.144	0.212	2.664
Takin-ADA	27.429	0.948	0.019	3.983	0.261	0.211	3.575

Table 2. Quantitative comparisons for self-reenactment and cross-reenactment methods.

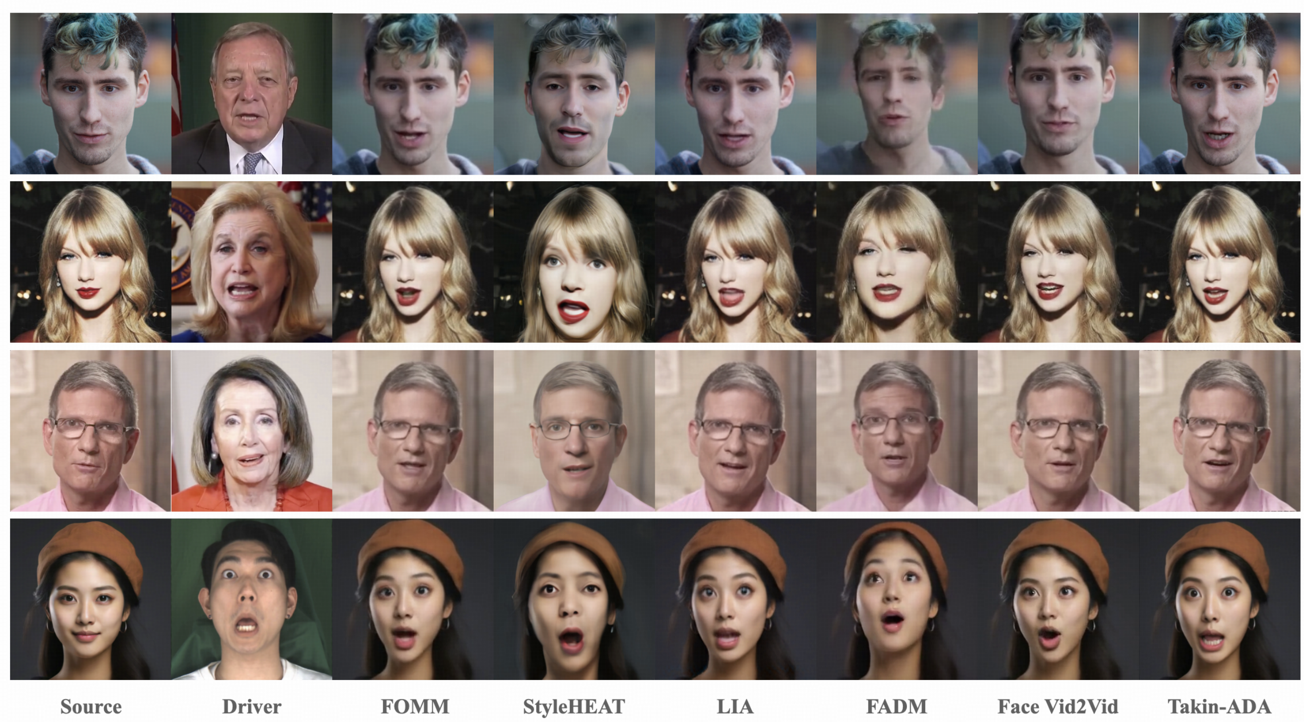

Figure 3. Qualitative comparisons of Cross-reenactment. This task involves transferring actions from a source portrait to a target portrait to evaluate each algorithm's ability to separate motion and appearance. The results highlight our method's superior ability in both motion transfer and appearance retention, while also excelling in the transfer of subtle micro-expressions and extreme facial expressions.

Cross-Reenactment (using another person's video for driving), the overall results tend to be lower compared to Self-Reenactment (using the current person's video). In Self-Reenactment, our algorithm achieved superior results for image structural metrics such as FID, CSIM, and LPIPS, validating the effectiveness of our motion representation in reconstructing images. Specifically, Takin-ADA achieved a FID score of 27.429, which is notably lower than FOMM and Vid2Vid, indicating a smaller distributional discrepancy between generated and real images. Additionally, the CSIM score of 0.937 surpasses other methods, demonstrating better identity preservation. The lowest LPIPS value of 0.019 further confirms the superior visual quality of our generated results. In the cross-reenactment task, our method also shows significant advantages, especially in terms of CSIM and LPIPS metrics. Our system effectively separates the driving actions and identity features, retaining the target head movements and expressions while preserving the source identity. The high MOS-ET score also reflects the high subjective satisfaction with our method. Takin-ADA achieved the best performance among all methods, with a CSIM score of 0.261 and a LPIPS score of 0.211. These results highlight our algorithm's outstanding ability to disentangle identity and motion when driving with different individuals, providing more natural, expressive, and high-fidelity facial animations.

Qualitative Results. Figure 3 presents a qualitative comparison of cross-reenactment methods. This task involves transferring actions from a source portrait to a target portrait to evaluate each algorithm's ability to separate mo-

Method	Subjective Evaluation			Objective Evaluation			
	MOS-R↑	MOS-N↑	MOS-LS↑	PSNR↑	SSIM↑	FID↓	LSE-D↓
MakeItTalk[62]	2.135	2.822	2.441	26.693	0.762	31.113	10.888
SadTalker[56]	3.783	2.148	3.573	26.105	0.753	32.539	7.748
AniPortrait[48]	3.529	2.329	3.474	25.172	0.731	33.434	7.968
AniTalker[28]	3.956	2.812	3.821	25.387	0.749	29.839	10.171
EDTalk[40]	2.943	3.152	3.752	26.978	**0.781**	28.043	**7.686**
Takin-ADA	**4.187**	**3.839**	**3.887**	**27.876**	0.779	**27.803**	7.764

Table 3. Quantitative comparisons with previous speech-driven methods.

tion and appearance. From the third row, it is clear that our method, Takin-ADA, excels in transferring subtle micro-expressions, effectively capturing and replicating delicate facial movements. From the fourth row, Takin-ADA also shows superior performance in handling extreme facial expressions, maintaining the integrity and authenticity of the facial features even under challenging conditions. These results highlight the robustness and effectiveness of Takin-ADA in both subtle and extreme expression transfer.

3.4. Audio-driven methods

We compare our method against leading speech-driven approaches, including MakeItTalk[62], SadTalker[56], AniPortrait[48], AniTalker[28] and EDTalk[40]. Table 3 presents the quantitative results of this comparison. Subjective evaluations consistently demonstrate that our method outperforms existing techniques in lip-sync accuracy(MOS-LS), naturalness(MOS-N), and Resolution(MOS-R), with particular emphasis on enhanced naturalness of movements. These improvements can be attributed to our sophisticated universal motion representation. Notably, our model demonstrates a superior ability to produce convincingly synchronized lip movements that accurately match the given phonetic sounds. Nevertheless, our SSIM[46] and LSE-D metric exhibits a slight decline compared to EDTalk, which we attribute to two primary factors: 1) EDTalk [40] is exclusively trained on lip movements, whereas our model predicts the full range of facial expressions. 2) the LSE-D metric emphasizes short-term alignment, 3) the metric is not utilized as a supervisory signal in our training process, thereby failing to sufficiently capture the long-term information essential for the comprehensibility of generated videos. This observation is further supported by the qualitative results presented in Figure 4, which underscore our model's capability to produce convincingly synchronized lip movements corresponding to the provided phonetic sounds.

Consistency with the longer pronunciation. Figure 4 demonstrates our model's proficiency in generating highly synchronized lip movements that correspond accurately to the given phonetic sounds. This visual representation underscores the model's capability to create realistic and precisely timed facial animations that align seamlessly with spoken language.

Figure 4. Visual comparison of the speech-driven method. Phonetic sounds are highlighted in red.

Emotion Control. Figure 5 presents a diverse array of our generated results, encompassing various emotional states. These examples vividly demonstrate our generation model's proficiency in interpreting emotional signals and producing talking face animations that closely correspond to the specified emotional parameters.

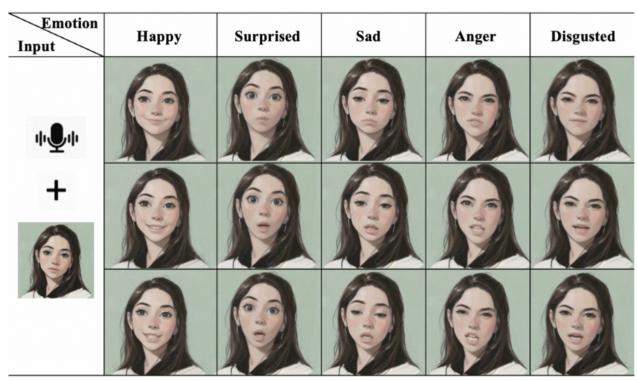

Figure 5. Generated results under different emotion offset (happy, surprised, sad, angry and disgusted, respectively).

The results unequivocally showcase the model's capacity to accurately capture and convey a wide spectrum of emotions through the generated facial expressions and movements. This underscores the system's effectiveness in translating emotional inputs into visually convincing and emotionally resonant animations.

3.5. Ablation Study

To further validate the effectiveness of our disentanglement between canonical and landmark information, we conducted an extensive ablation study using various methods. First, to evaluate the performance of our model without the canonical loss ($\mathcal{L}_{canonical}$), we observed the resulting metrics and compared them against a fine-tuned vid2vid baseline. This comparison, detailed in Table 4, demonstrates significant improvements across all metrics when either component is added. The exclusion of $\mathcal{L}_{canonical}$ resulted in moderate improvements, with an FID of 27.429, CSIM of 0.948, MOS-ET of 3.983, and PSNR of 24.663. The exclusion of $\mathcal{L}_{landmark}$ yielded better results, achieving an FID of 61.1, CSIM of 0.69, MOS-ET of 3.6, and PSNR of 29.6. By incorporating both $\mathcal{L}_{canonical}$ and $\mathcal{L}_{landmark}$, our complete method achieved the best results. These results highlight the powerful synergy of these disentanglement losses, leading to enhancements in image quality, structural similarity, and expression transfer. Our findings emphasize the importance of these components in ensuring the motion encoder effectively focuses on relevant motion-related information, thereby improving the overall performance of our approach. This analysis is comprehensively demonstrated in Table 2, reinforcing the significance of disentanglement methods in achieving superior image re-enactment quality.

Method	FID↓	CSIM↑	MOS-ET↑	PSNR↑
Face Vid2Vid fine-tuned	28.444	0.945	3.451	19.235
Ours w/o $\mathcal{L}_{\text{canonical}}$	28.721	0.947	3.542	22.254
Ours w/o $\mathcal{L}_{\text{landmark}}$	27.828	0.948	3.662	23.619
Ours	**27.429**	**0.948**	**3.983**	**24.663**

Table 4. Quantitative comparisons of disentanglement methods in Self-Reenactment setting

4. CONCLUSIONS

In this paper, we introduced Takin-ADA, an innovative two-stage framework for real-time audio-driven animation of single-image portraits with controllable expressions using 3D implicit keypoints. Our approach addresses critical limitations in existing methods, such as expression leakage, subtle expression transfer, and audio-driven precision. By employing a canonical loss and a landmark-guided loss to enhance the transfer of subtle expressions while simultaneously mitigating expression leakage in the first stage, and a state-of-the-art audio-conditioned diffusion model based on HuBERT features in the second stage, Takin-ADA achieves high-resolution (512×512) facial animations at up to 42 FPS on an RTX 4090 GPU. Our extensive evaluations demonstrate that Takin-ADA consistently outperforms existing solutions in video quality, facial dynamics realism, and naturalness of head movements.

While Takin-ADA shows significant advancements, it has some limitations, including minor inconsistencies in complex backgrounds and edge blurring during extreme facial shifts. Future work will focus on improving the temporal coherence and rendering quality of the framework. Takin-ADA sets a new benchmark in single-image portrait animation, opening new possibilities for applications like virtual hosts, online education, and digital human interactions, and providing a robust foundation for future research in this evolving field.

Acknowledgement

An acknowledgement is used to thank the person, fund, etc., that support this work.

References

[1] 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*, 142:104911, 2024. 3

[2] E. Burkov, I. Pasechnik, A. Grigorev, and V. Lempitsky. Neural head reenactment with latent pose descriptors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13786–13795, 2020. 3

[3] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 6

[4] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu. Lip movements generation at a glance. In *Proceedings of the European conference on computer vision (ECCV)*, pages 520–535, 2018. 3

[5] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520, 2017. 5

[6] Q. Chen, Z. Ma, T. Liu, X. Tan, Q. Lu, K. Yu, and X. Chen. Improving few-shot learning for talking face system with tts data augmentation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 1

[7] Z. Chen, J. Cao, Z. Chen, Y. Li, and C. Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions, 2024. 6

[8] Z. Dai. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019. 6

[9] R. Danecek, M. J. Black, and T. Bolkart. Emoca: Emotion driven monocular face capture and animation, 2022. 1

[10] N. Drobyshev, J. Chelishev, T. Khakhulin, A. Ivakhnenko, V. Lempitsky, and E. Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2663–2671, 2022. 3

[11] 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*, pages 4281–4289, 2023. 6

[12] C. Du, Y. Guo, F. Shen, Z. Liu, Z. Liang, X. Chen, S. Wang, H. Zhang, and K. Yu. Unicats: A unified context-aware text-to-speech framework with contextual vq-diffusion and vocoding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17924–17932, 2024. 3

[13] 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 (CVPR)*, pages 18770–18780, June 2022. 1

[14] Y. Feng, H. Feng, M. J. Black, and T. Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Trans. Graph.*, 40(4), July 2021. 1

[15] 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*, pages 5609–5619, 2023. 2

[16] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020. 5, 6

[17] J. Guo, D. Zhang, X. Liu, Z. Zhong, Y. Zhang, P. Wan, and D. Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control, 2024. 1, 5

[18] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3

[19] T. He, J. Guo, R. Yu, Y. Wang, J. Zhu, K. An, L. Li, X. Tan, C. Wang, H. Hu, et al. Gaia: Zero-shot talking avatar generation. *arXiv preprint arXiv:2311.15230*, 2023. 3

[20] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[21] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 5

[22] F.-T. Hong and D. Xu. Implicit identity representation conditioned memory compensation network for talking head video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23062–23072, 2023. 3

[23] F.-T. Hong, L. Zhang, L. Shen, and D. Xu. Depth-aware generative adversarial network for talking head video generation, 2022. 1

[24] F.-T. Hong, L. Zhang, L. Shen, and D. Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3397–3406, 2022. 3

[25] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. 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. 6

[26] J. Jiang, C. Liang, J. Yang, G. Lin, T. Zhong, and Y. Zheng. Loopy: Taming audio-driven portrait avatar with long-term motion dependency. *arXiv preprint arXiv:2409.02634*, 2024. 3

[27] 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*, pages 3387–3396, 2022. 3

[28] T. Liu, F. Chen, S. Fan, C. Du, Q. Chen, X. Chen, and K. Yu. Anitalker: Animate vivid and diverse talking faces through identity-decoupled facial motion encoding, 2024. 1, 6, 8

[29] S. Ma, T. Simon, J. Saragih, D. Wang, Y. Li, F. D. La Torre, and Y. Sheikh. Pixel codec avatars. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 64–73, 2021. 1

[30] Y. Ma, S. Zhang, J. Wang, X. Wang, Y. Zhang, and Z. Deng. Dreamtalk: When emotional talking head generation meets diffusion probabilistic models, 2024. 1

[31] A. Mallya, T.-C. Wang, and M.-Y. Liu. Implicit warping for animation with image sets. *Advances in Neural Information Processing Systems*, 35:22438–22450, 2022. 3

[32] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017. 6

[33] Z. Peng, H. Wu, Z. Song, H. Xu, X. Zhu, J. He, H. Liu, and Z. Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20687–20697, 2023. 1

[34] K. R. 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*, MM '20, page 484–492, New York, NY, USA, 2020. Association for Computing Machinery. 1, 3, 6

[35] Y. Ren, G. Li, Y. Chen, T. H. Li, and S. Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13759–13768, 2021. 2

[36] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019. 2, 3, 7

[37] A. Siarohin, O. J. Woodford, J. Ren, M. Chai, and S. Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662, 2021. 3

[38] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

[39] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from

audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 3

[40] S. Tan, B. Ji, M. Bi, and Y. Pan. Edtalk: Efficient disentanglement for emotional talking head synthesis, 2024. 1, 6, 8

[41] L. Tian, Q. Wang, B. Zhang, and L. Bo. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*, 2024. 3

[42] 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*, pages 700–717. Springer, 2020. 6

[43] S. Wang, L. Li, Y. Ding, and X. Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *International Joint Conferences on Artificial Intelligence Organization*, 2021. 1

[44] T.-C. Wang, A. Mallya, and M.-Y. Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10034–10044, 2021. 1, 3, 4, 5, 6, 7

[45] Y. Wang, D. Yang, F. Bremond, and A. Dantcheva. Latent image animator: Learning to animate images via latent space navigation. In *International Conference on Learning Representations*, 2022. 1, 7

[46] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6, 8

[47] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[48] H. Wei, Z. Yang, and Z. Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation, 2024. 8

[49] S. Xu, G. Chen, Y.-X. Guo, J. Yang, C. Li, Z. Zang, Y. Zhang, X. Tong, and B. Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time, 2024. 1

[50] F. Yin, Y. Zhang, X. Cun, M. Cao, Y. Fan, X. Wang, Q. Bai, B. Wu, J. Wang, and Y. Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European conference on computer vision*, pages 85–101. Springer, 2022. 3, 7

[51] Z. Yu, Z. Yin, D. Zhou, D. Wang, F. Wong, and B. Wang. Talking head generation with probabilistic audio-to-visual diffusion priors, 2022. 1, 3

[52] E. Zakharov, A. Ivakhnenko, A. Shysheya, and V. Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 524–540. Springer, 2020. 2

[53] B. Zeng, X. Liu, S. Gao, B. Liu, H. Li, J. Liu, and B. Zhang. Face animation with an attribute-guided diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 628–637, 2023. 7

[54] B. Zhang, C. Qi, P. Zhang, B. Zhang, H. Wu, D. Chen, Q. Chen, Y. Wang, and F. Wen. Metaportrait: Identity-preserving talking head generation with fast personalized adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22096–22105, 2023. 2, 6

[55] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[56] 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*, pages 8652–8661, 2023. 3, 6, 8

[57] 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*, pages 3661–3670, 2021. 1, 6

[58] J. Zhao and H. Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022. 3

[59] 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, 2023. 1

[60] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4176–4186, 2021. 1, 3

[61] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6):1–15, 2020. 1

[62] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics*, 39(6):1–15, Nov. 2020. 1, 6, 8