This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SynchroRaMa : Lip-Synchronized and Emotion-Aware Talking Face Generation via Multi-Modal Emotion Embedding

Phyo Thet Yee
IIT Ropar, India
phyo.22csz0009@iitrpr.ac.in

Dimitrios Kollias
Queen Mary University of London, UK
d.kollias@qmul.ac.uk

Sudeepta Mishra
IIT Ropar, India
sudeepta@iitrpr.ac.in

Abhinav Dhall
Monash University, Australia
abhinav.dhall@monash.edu

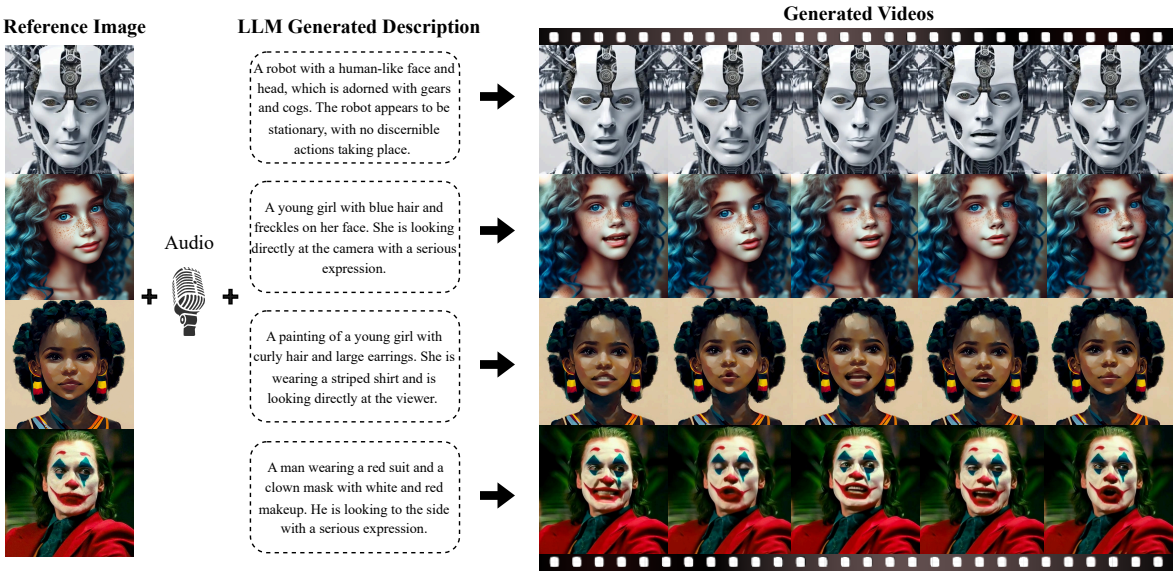


Figure 1. We propose SynchroRaMa, an expressive talking face generation framework. Given a single reference image, audio, and a textual description, our model can generate talking face videos featuring lip-synchronized, expressive facial expressions and emotional cues while maintaining identity consistency. Please see more results in the supplementary material.

Abstract

Audio-driven talking face generation has received growing interest, particularly for applications requiring expressive and natural human-avatar interaction. However, most existing emotion-aware methods rely on a single modality (either audio or image) for emotion embedding, limiting their ability to capture nuanced affective cues. Additionally, most methods condition on a single reference image, restricting the model's ability to represent dynamic changes in actions or attributes across time. To address these issues, we introduce SynchroRaMa, a novel framework that integrates a multi-modal emotion embedding by combining emotional signals from text (via sentiment analysis) and audio (via speech-based emotion recognition and audio-derived valence-arousal features), enabling the generation of talking face videos with richer and more authentic emotional expressiveness and fidelity. To ensure natural head motion and accurate lip synchronization, SynchroRaMa includes an audio-to-motion (A2M) module that generates motion frames aligned with the input audio. Finally, SynchroRaMa incorporates scene descriptions generated by Large Language Model (LLM) as additional textual input, enabling it to capture dynamic actions and high-level semantic attributes. Conditioning the model on both visual and textual cues enhances temporal consistency and visual realism. Quantitative and qualitative experiments on benchmark datasets demonstrate that SynchroRaMa outperforms the state-of-the-art, achieving improvements in image quality, expression preservation, and motion realism. A user study further confirms that SynchroRaMa achieves higher subjective ratings than competing methods in overall naturalness, motion diversity, and video smoothness. Our project page is available at https://novicemm.github.io/synchrorama.

1. Introduction

Talking face generation [1, 6, 7, 18, 33, 36] aims to animate a portrait image by integrating audio. It has gained popularity in various domains such as video games, film industries, social media, digital marketing and education sectors. Existing approaches to talking face generation primarily utilize either GANs [8] or diffusion models [5, 11]. GAN-based methods [14, 20] use a combination of audio and visual encoders to extract features from speech and video frames, which are then processed by a generator network to produce synchronized lip movements. In contrast, diffusion-based models generate results through an iterative refinement process, leading to higher-quality and more temporally coherent outputs. Despite these improvements, generating a realistic talking face remains challenging, as it requires precise lip synchronization, and head poses with the given speech. In addition to lip synchronization, maintaining the visual coherence and capturing the richness of expression and emotion remain open challenges in talking face generation.

To address these limitations, we introduce SynchroRaMa, a novel framework designed to generate high-quality, emotionally expressive, and lip-synchronized talking faces from the audio input, textual description and a reference image. Most previous works on emotion-aware talking face generation primarily focus on a single modality, such as text, audio or visual cues for emotion embedding. However, relying on a single modality limits model performance, as each modality contains its specific constraints, and individually, none can fulfill all requirements perfectly. Therefore, our work leverages multi-modal emotion embedding by combining textual sentiment analysis, speech-based emotion recognition, and valence-arousal (VA) emotion embedding derived from audio signals.

Diffusion-based talking face generation approaches typically use a reference network and denoising UNet as their backbone. Visual appearance information is fed into Reference network, which is then integrates with the Denoising UNet during the denoising process. However, providing only visual information is insufficient, as existing methods mostly use a single frame extracted from the input video as the reference image during training. Relying on a single frame to represent the entire video may fail to capture potential changes in scenes, subject's actions, and attributes in subsequent frames. As a result, the generated video may lack detailed appearance consistency. To overcome this, we incorporate textual descriptions that contain changes in scenes (temporal) information as additional input. Recent advancements in visual large language models such as VideoLLaMA2 [2], can generate comprehensive textual descriptions of the entire video, capturing changes in scenes, actions and attributes. This textual information complements the detailed visual cues of the extracted reference image, enhancing the visual quality of the generated video.

Another important aspect in talking face generation is motion consistency. Generated video should exhibit realistic head movement and lip should be synchronized with the speech. To ensure this, we also introduce an audio-to-motion module, producing motion frames driven by the audio. Training the model with these audio-driven motion frames guarantees realistic head movements and accurate lip synchronization. Furthermore, we introduce several loss functions in our work, including syncloss, emo loss, facial action units (AU) loss, and attr-action loss.

The **main contributions** of our paper can be summarized as: 1) We propose a novel talking face generation framework, which leverages multi-modal information, including visual, textual and audio data. 2) We introduce a multi-modal emotion embedding module to enrich emotional expressiveness in the generated videos. 3) We embed LLM generated scene description to make the generation better aligned to context. 4) We present an Audio-to-Motion (A2M) module, designed to generate realistic motion frames synchronized with the audio. Our quantitative and qualitative experiments demonstrate that our approach achieves superior performance compared to the state-of-the-art (SOTA) methods.

2. Related Work

2.1. Diffusion-based Talking Face Generation

Diffusion models have recently gained significant attention in talking face generation. Hallo [36] leverages the Stable Diffusion Model [25] with the ReferenceNet to maintain appearance consistency and introduces a hierarchical audio-visual cross-attention mechanism to align audio and visual features. VASA-1 [37] operates in a disentangled latent space to enable precise and expressive facial animations. AniTalker [15] use universal motion representation to capture a wide range of facial dynamics including subtle expressions and head movements. X-Portrait [35] employs a conditional diffusion model enhanced with a hierarchical patch-based local control module for accurate and coherent motion transfer. Diff2Lip [19] and DiffTalk [26] use Latent Diffusion Models (LDMs) conditioned on audio features, reference images, masked ground-truth images, and facial landmarks. GAIA [10] disentangles motion and appearance to preserve identity while enabling speech-driven motion synthesis. EchoMimic [1] and AniPortrait [33] employ concurrent training, conditioning simultaneously on audio signals and facial landmarks to generate realistic portrait animations. VividTalk [27] leverages a 3D hybrid prior to decompose facial expressions and head movements, using a learnable codebook for natural motion and a dual-branch Motion VAE to generate dense motion fields. EMO [28] utilizes Audio2Video diffusion model, integrating weak condi-

tion constraints such as face locators and motion guidance, bypassing the need for intermediate 3D models. MODA [16] employs a mapping-once network with dual attention mechanisms to convert audio into motion representations, where one attention captures accurate lip-sync and the other models natural head and eye movements.

3. Method

3.1. Model Architecture

Given a reference image, textual description, and input audio, our model generates a lip-synchronized talking face video while preserving natural head movements, facial expressions, emotions, and overall appearance consistency. As shown in Figure 2, our framework integrates several critical components: Denoising UNet [25] as the backbone network, ReferenceNet [12] to encode the reference image, proposed Audio-to-Motion (A2M) module to generate the motion frames synchronized with the input audio, and proposed multi-modal emotion embedding module to align the generated video with the emotional content of the audio. The careful incorporation of these components ensures the generation of realistic and contextually coherent talking face videos. The following sections present a detailed explanation of each component.

3.2. Denoising UNet

In our framework, we adopt the Denoising UNet architecture from the Stable Diffusion [25] 1.5 as the backbone network by incorporating 4 distinct attention mechanisms within each Transformer block (Figure 2(a)). Specifically, Spatial Attention layer focuses on the important spatial regions, ensuring the model captures the relevant facial features such as mouth, eyes, and expressions while preserving the identity consistency. Audio Attention layer integrates the audio features into the generation process. Cross Attention layer bridges the multiple modalities such as reference image, audio, and textual description by dynamically aligning them throughout the denoising process. Finally, Temporal Attention layer ensures smooth and coherent motion transitions between the consecutive video frames. These integrated attention mechanisms enhance the model's ability to capture and integrate spatial and temporal relationships effectively within the generation process.

3.3. ReferenceNet

ReferenceNet employs the same architecture as the Denoising UNet and guides the generation process by embedding both reference image and the textual input (Figure 2(a)). It ensures the preservation of facial identity and background consistency. Within the ReferenceNet, each Transformer block employs a Spatial Attention mechanism to extract features from the reference image. These features are then passed to the Denoising UNet via the corresponding Spatial Attention layers. Additionally, Cross Attention layer integrates the textual description that contains attributes and actions of the corresponding video as additional conditioning input. This provides semantic guidance to further enhance appearance consistency. This integration of visual and textual conditioning ensures identity and background consistency, as well as high-quality video generation.

3.4. Video Description

Previous works [1, 33, 36] use one extracted frame from the original video as a reference image during training to provide appearance information (e.g., identity and background details). However, relying on a single image may fail to capture dynamic changes in actions and attributes that occur over time, e.g., a person sitting in the first frame might change position in subsequent frames - a single image cannot reflect these changes. To address this limitation, we incorporate a textual description that represents the entire video alongside the reference image. Specifically, we employ VideoLLaMA2 [2] to generate a video description that captures fine-grained details such as attributes (e.g., "wearing earrings") and actions (e.g., "turning head"). In [40, 41], CLIP embeddings are used to maintain the identity and facial expressions. Inspired by these works, we use a CLIP text encoder to encode the VideoLLaMA2 text descriptions into rich semantic features, and feed them into the model through cross-attention layers in the ReferenceNet and the Denoising UNet (Figure 2(a)). During training, ReferenceNet processes the reference image and textual description to extract visual-textual appearance features, which are integrated into the Denoising UNet, then fused with emotion and audio features. This conditioning strategy ensures that high-level semantic cues from text influence both attributes and actions throughout the video.

3.5. Audio-to-Motion (A2M) Module

The main challenge in audio-driven motion generation is to generate temporally consistent and expressive motion sequences that align with the input audio. To address this, we propose a VAE-based Audio-to-Motion module, which takes both a reference image and audio features as input and generates motion frames (Figure 2(b)). We first extract audio features using wav2vec 2.0, and use them as conditional inputs to the VAE encoder. In VAE, the encoder typically outputs parameters (μ, σ) that define a Gaussian distribution, from which a latent variable z_q is sampled as $z_q \sim \mathcal{N}(0, 1)$. However, this Gaussian prior tends to force the posterior toward the mean, which limits output diversity and generative power. To overcome this, we follow [24, 38, 39] and introduce a Volume-Preserving Normalizing Flow (VP-Flow) that transforms the simple latent variable z_q into a more expressive latent representation z_p (i.e, $z_p = f_{VP-Flow}(z_q)$) (Figure 2(c)). The VP-Flow is

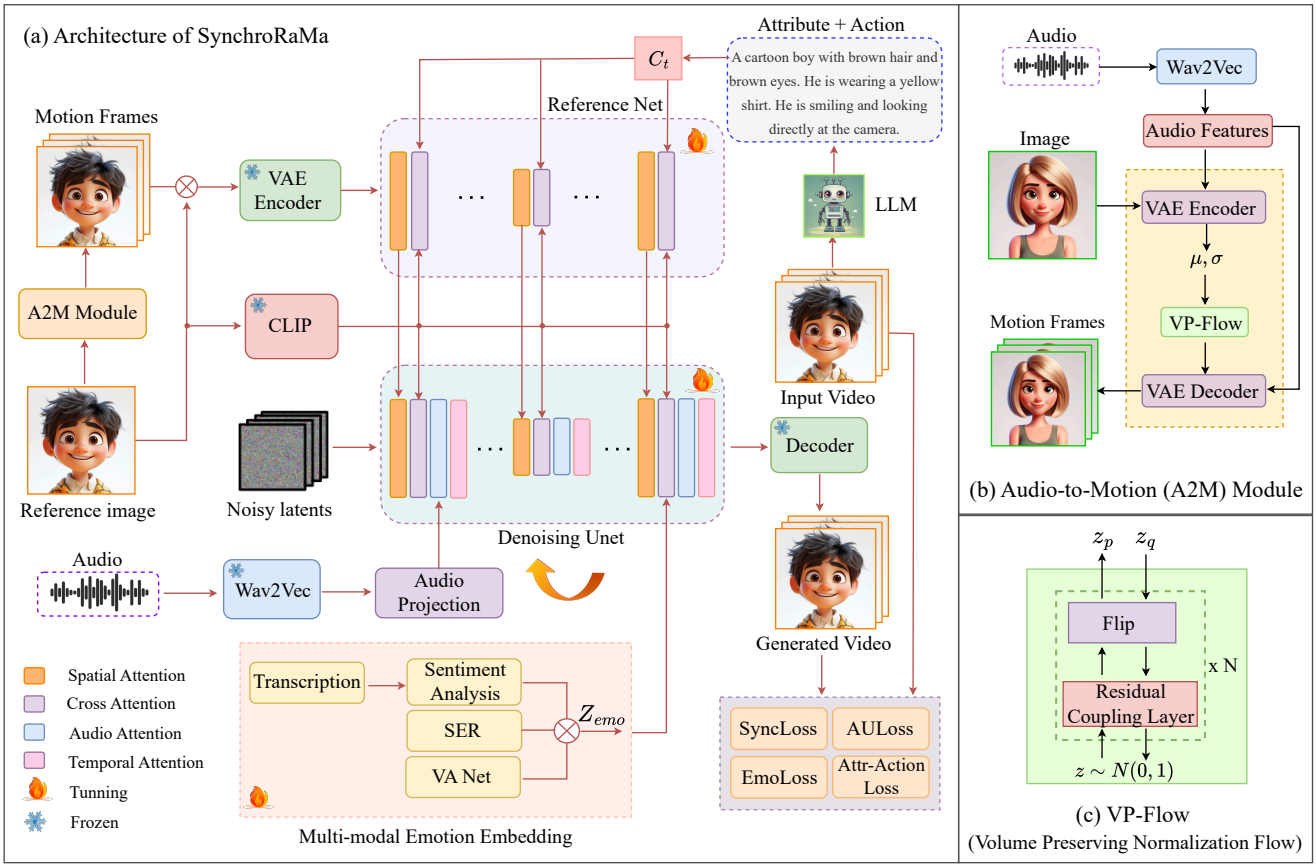


Figure 2. (a) Overall architecture of SynchroRaMa. The framework takes a reference image, LLM-generated textual description, and driving audio as input. It operates in two stages. In the first stage, ReferenceNet encodes visual-textual appearance features from the reference image and textual description. Motion frames are also added to ensure temporal smoothness and coherence across frames. In the second stage, these features are integrated with audio projection and a multi-modal emotion embedding within the Denoising UNet during the diffusion process. Motion frames are generated by (b). C_t means CLIP Text Encoder. (b) Architecture of the Audio-to-Motion (A2M) Module. Audio features are extracted using Wav2Vec 2.0, then encoded by a VP-Flow-based VAE, which consists of residual coupling layers and flip operations. (c) Architecture of VP-Flow.

composed of a stack of residual affine coupling layers and channel-wise flip operations, which ensure invertibility and maintain the volume of the latent space during transformation. We then pass the transformed latent variable z_p, together with the audio features to the VAE decoder, which then synthesizes a sequence of visual motion frames that are aligned with the input audio. Therefore, VP-Flow-based conditioning in our A2M module enhances the diversity and realism of generated motion, while the audio features ensure accurate lip sync and smooth transitions across frames.

3.6. Multi-modal Emotion Embedding Module

Emotion embedding is an important aspect in talking face generation to provide more realistic and expressive facial movements aligned with the driven audio. To achieve this, we introduce a multi-modal emotion embedding mechanism that integrates emotional cues from three modalities - text, audio, and valence-arousal (VA) features (Figure 2(a)), providing a richer perspective than any single modality alone. First, we transcribe the input audio using Whisper [23] and perform sentiment analysis on the resulting text using emotion-english-distilroberta [9]. Second, we perform speech emotion recognition (SER) on the audio using [22]. Third, we extract valence and arousal features from the audio using a fine-tuned wav2vec 2.0 [29]. To ensure accurate extraction of valence-arousal features, we remove background music and divide the audio into 50% overlapping segments to capture temporal variations. The features from each segment are then concatenated to form the final representation. Let E_t, E_{ser}, and E_{va} denote emotion embeddings from textual sentiment analysis, SER, and VA features, respectively, computed as $E_t = f_{\text{Transcription}}(a)$, $E_{ser} = f_{\text{SER}}(a)$, $E_{va} = f_{\text{VA}}(a)$. These embeddings are concatenated to form the final emotion embedding:

$$E = \text{Concat}(E_t, E_{ser}, E_{va}) \quad (1)$$

This combined embedding is fed into the denoising UNet via a cross-attention layer. By employing multiple modal-

ities, our method captures a broader range of emotional nuances, resulting in a more expressive talking face that closely aligns with the speaker's intended emotion.

3.7. Loss Functions

Alongside standard diffusion loss, we introduce auxiliary losses to ensure that the generated video maintains temporal coherence, emotional expressiveness, identity consistency, and semantic alignment with the input modalities.

Sync Loss. This loss enforces precise temporal alignment between generated lip movements and input audio, which is critical for perceptual realism in talking face. We evaluate synchronization using [13], which estimates the temporal offset between audio and video, and detects the location of misalignment. The loss compares the global temporal alignment signals, which are more robust to the local jitter and better reflect the human perception. It is defined as:

$$\mathcal{L}_{\text{sync}} = (\Delta t_p - \Delta t_{gt})^2 + (t_p - t_{gt})^2 \tag{2}$$

where: Δt_p and Δt_{gt} are the predicted and ground truth temporal offset magnitudes between the audio and generated video, and t_p and t_{gt} are corresponding timestamps where misalignment occurs.

Emo Loss. This loss enhances emotionally consistent facial expressions by aligning the temporal emotional dynamics of the generated video with those of the ground truth. Specifically, it encourages the output to reflect similar variations in valence and arousal over time. To compute it, we first remove the background music and divide the audio (from both generated and ground truth videos) into 50% overlapping segments. Employing overlapping segments ensures smoother emotional transitions and reduces sudden shifts or potential artifacts that may occur from non-overlapping segments. For each segment, we extract valence and arousal values using a transformer-based model [29], a fine-tuned variant of wav2vec 2.0 that predicts continuous emotional dimensions (valence and arousal) directly from the audio. This yields a sequence of valence and arousal pairs that captures the evolving emotional state throughout the video. The loss provides fine-grained, temporally-aware supervision, enhancing the emotional expressiveness and coherence of the generated facial behavior. The loss is defined as:

$$\mathcal{L}_{\text{emo}} = \frac{1}{K} \sum_{k=1}^{K} \left[\left(v_p^{(k)} - v_{gt}^{(k)} \right)^2 + \left(a_p^{(k)} - a_{gt}^{(k)} \right)^2 \right] \tag{3}$$

where: K is the number of audio segments; $v_p^{(k)}$ and $a_p^{(k)}$ & $v_{gt}^{(k)}$ and $a_{gt}^{(k)}$ are the VA values of the audio of generated video and of ground truth video for segment k.

Facial Action Unit (AU) Loss. This loss provides fine-grained supervision to enhance the expressiveness and realism of the generated videos. It enables the model to accurately capture subtle facial muscle movements, such as eyebrow raises or lip stretches, which are often overlooked by global emotion descriptors such as valence-arousal values. This targeted supervision helps improve the semantic accuracy and coherence of facial expressions, ensuring that localized facial actions align with the intended emotion. Furthermore, AU loss promotes the temporal and spatial consistency across video frames. It also complements the multi-modal emotion embedding by refining the local expression details, while the embedding captures the overall emotional tone. AU loss is computed as the squared L2 distance between the predicted AUs of the generated video and the ground truth AUs from the original video. Formally, it is defined as:

$$\mathcal{L}_{AU} = \frac{1}{T \cdot N} \sum_{t=1}^{T} \sum_{i=1}^{N} \left(AU_{p,i}^{(t)} - AU_{gt,i}^{(t)} \right)^2 \tag{4}$$

where: T is the number of frames; N is the number of AUs; $AU_{p,i}^{(t)}$ is the i-th AU prediction at frame t; $AU_{gt,i}^{(t)}$ is i-th AU ground truth at frame t.

Attr-Action Loss This loss ensures that the generated video preserves the high-level semantic attributes and actions described in the original video, e.g., head movements, or appearance cues. We generate the textual descriptions for both predicted and ground truth videos using VideoLLaMA2 [2], and compute the loss as the cosine distance between their textual embeddings:

$$\mathcal{L}_{\text{attr-action}} = 1 - \cos\left(\mathbf{e}_p, \mathbf{e}_{gt}\right) \tag{5}$$

where: $\mathbf{e}_p$ and $\mathbf{e}_{gt}$ are the textual embeddings of the predicted and ground truth video, respectively; $\cos(\cdot, \cdot)$ denotes the cosine similarity between the two embedding vectors.

4. Experiments

Implementation Details. We conduct all experiments on NVIDIA A100 GPUs. The model is trained in two stages, each for 30k steps, with batch size 4 and resolution 512x512. In the first stage, we train the model to encode appearance information using a reference image and its corresponding textual description as inputs to the ReferenceNet. Each training sample consists of a 14-frame video clip, from which one frame is randomly selected as the reference frame and another as the target frame. Additionally, motion frames are added to ensure temporal smoothness and coherence across frames. During this stage, the VAE encoder/decoder and the CLIP image/text encoders are kept frozen, while only the ReferenceNet and the Denoising UNet are optimized. Both networks are initialized with weights from the original Stable Diffusion model. In the second stage, the model is trained on full video sequences with audio injection for audio-visual alignment and emotion embedding to capture emotional cues. In both stages, a

Table 1. Comparison with SOTA approaches on the HDTF [43] (top) and MEAD [31] (bottom) datasets.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓	E-FID ↓	F1 ↑	Sync ↑	CCC_V ↑	CCC_A ↑
Hallo	30.63	0.71	0.18	28.91	156.32	1.38	0.68	7.58	0.53	0.56
EchoMimic	28.90	0.51	0.44	71.33	153.84	1.32	0.63	6.20	0.51	0.49
VExpress	29.41	0.61	0.36	58.63	350.18	1.85	0.46	**8.07**	0.32	0.30
Aniportrait	30.64	0.66	0.23	42.38	449.51	1.97	0.41	3.15	0.11	0.12
Ours	**32.97**	**0.73**	**0.17**	**27.67**	**149.67**	**1.24**	**0.71**	7.03	**0.56**	**0.58**
Hallo	31.69	**0.86**	0.11	31.30	**147.27**	1.41	0.65	6.25	0.54	0.51
EchoMimic	29.07	0.68	0.42	63.25	164.75	1.50	0.61	5.98	0.53	0.50
VExpress	29.34	0.65	0.39	61.48	443.96	2.31	0.42	**7.25**	0.34	0.25
Aniportrait	30.36	0.80	0.16	51.40	508.36	2.87	0.31	2.06	0.13	0.11
Ours	**32.21**	**0.86**	**0.09**	**28.47**	147.87	**1.36**	**0.67**	6.84	**0.57**	**0.54**

learning rate of $1e^{-5}$ is used. During inference, the model takes a reference image, its textual prompt, and driving audio as inputs, and generates a video sequence by animating the reference image in sync with the audio. We use DDIM sampling with 40 steps to generate each output video clip.

Datasets. We train our model using VFHQ [34], HDTF [43], and a selection of in-the-wild scene clips shared by Hallo3 [3]. To enhance training data quality: all videos are resized to 512x512 resolution; we filter out extreme side-profile views by detecting facial landmarks with MediaPipe [17] (ensuring clarity in the lip region); videos containing multiple speakers are excluded (maintaining speaker identity consistency). In the end, we obtain around 80 hours of video data, with individual clips ranging from 3 to 20 seconds in length. All videos are standardized to 25 fps to ensure temporal consistency across samples. Audio tracks are resampled to 16 kHz and normalized. When present, background music and noise are removed, as they can negatively impact emotion recognition and audio-to-motion alignment. For evaluation, we use 100 videos from each of HDTF [43] and emotion-aware MEAD [31].

Evaluation Metrics. We employ several metrics to evaluate the performance of our model. To assess the similarity between the generated and ground truth images, we compute PSNR, SSIM [32], and LPIPS [42]. Additionally, we use FID and FVD to measure how closely the generated data matches that of the actual data. To evaluate the accuracy of facial expressions, we compute: (i) Expression-FID (E-FID) for AUs (following [1, 28] - expression parameters are extracted using face reconstruction model [4]); (ii) F1 score for AUs; (iii) Concordance Correlation Coefficient (CCC) for Valence (CCC_V) and Arousal (CCC_A). Following [16], we adopt the confidence score from SyncNet (Sync) [21] to measure audio-visual synchronization.

5. Comparison with state-of-the-art Methods

Quantitative Comparison. We perform quantitative comparison with several SOTA, Hallo [36], Aniportrait [33], Echomimic [1] and VExpress [30] on the HDTF and MEAD datasets. We evaluate all methods using their publicly available checkpoints. As shown in Table 1, our approach outperforms existing methods across all metrics, except for sync confidence on the HDTF and MEAD datasets, and FVD on MEAD. In terms of image-level quality, our method achieves better FID, PSNR, SSIM, and LPIPS scores compared to the SOTA approaches. According to E-FID, F1 scores of AUs, CCC_V and CCC_A, our method outperforms competing methods in both expression and emotion preservation. Better performance across these metrics indicates that the generated face videos contain appropriate expressions and emotions, which are reflected in the

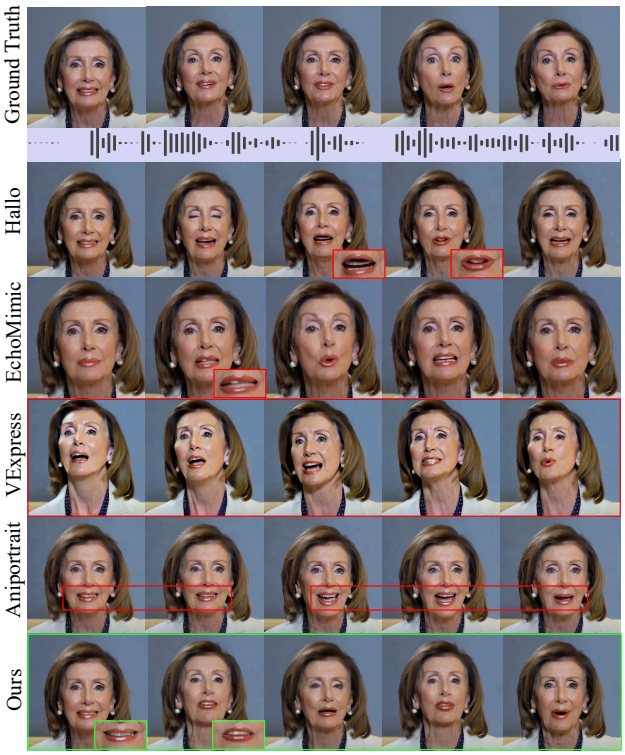


Figure 3. Qualitative comparison with the SOTA approaches on HDTF [43]. Please zoom in for a better view.



Figure 4. Qualitative comparison with the SOTA approaches on MEAD [31]. Please zoom in for a better view.

input audio and reference image. The lower FVD score demonstrates that our approach achieves better video quality. Additionally, our lip-synchronization performance is comparable to that of VExpress and Hallo. Notably, despite using less training data than the other methods, our approach achieves better results across various metrics.

Qualitative Comparison. We perform a qualitative comparison of our proposed method with SOTA approaches [1, 30, 33, 36] on both the HDTF and MEAD datasets. To evaluate the comparison, we extract frames from the original videos and use the first frame as the reference image. The corresponding audio is also extracted and use it as the driving audio. Hallo, Echomimic, and VExpress achieve accurate lip sync on both datasets. However, in Figure 3, Hallo produces artifacts in some frames (e.g., a black region around the teeth and unrealistic teeth), while Echomimic exhibits inconsistent motion between frames and unnatural lip (e.g., teeth appearing on the lips). VExpress often fail to generate the correct pose (Figure 3, 4), does not maintain identity (Figure 3) and introduces excessive motion. Aniportrait struggles with lip sync accuracy (Figure 3, 4), showing minimal or indistinct lip movements across frames. In contrast, our approach generates talking face videos with good overall quality, natural lip movements, and consistent identity. Furthermore, by incorporating emotional awareness in addition to lip synchronization, our model generates videos that are both realistic and emotionally expressive. As shown in Figure 4, our method generates emotions that appear more realistic and better aligned with the ground truth than those generated by SOTA approaches.

User Study. We conduct a user study to further evaluate the quality of videos generated by our method and by other SOTA methods [1, 30, 33, 36]. All videos in the study include a balanced representation of genders, with varied ages, poses, and expressions. The study contained 20 participants, all of whom are Master's or PhD students with a background in Computer Science. Each participant is shown videos generated from the same image and audio inputs across all methods. They are then asked to rate each video on a scale from 1 to 5 based on lip sync, motion diversity, video smoothness, and overall naturalness. We collect the ratings and compute the average percentage score for each method. The results are presented in Table 2. Participants give higher ratings to our approach in terms of video quality, naturalness and motion consistency. Our performance in lip sync is comparable to that of other methods, which is consistent with the findings presented in Table 1.

Table 2. Results (in %) of the user study for comparison of our method with SOTA methods.

Method	Lip Sync.	Motion Diversity	Video Smoothness	Overall Naturalness
Hallo	25.73	23.61	28.53	27.35
EchoMimic	21.11	19.02	10.82	20.30
VExpress	**26.10**	7.25	5.23	7.58
Aniportrait	2.05	5.37	2.55	2.00
Ours	25.01	**44.75**	**52.87**	**42.77**

6. Ablation Studies

We perform the following ablation studies to evaluate contribution of different components of our method:

Table 3. Effect of addition of Audio-to-Motion (A2M) Module.

Method	Sync ↑	FVD ↓
w/o A2M module	4.33	182.25
w/ A2M module	**6.84**	**147.87**

Table 4. Effect of incldluing LLM generated textual description.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
w/o textual integration	31.30	0.62	0.15
w/ textual integration	**32.21**	**0.86**	**0.09**

Multi-modal Emotion Embedding Module. To evaluate the effectiveness of multi-modal emotion embedding module, we perform an experiment (Table 5) by excluding this component. Without the module, the model fails to generate appropriate emotional facial expressions that correspond to audio cues (Figure 5, top-left). In contrast, when it is included, the model accurately captures and generates the intended emotional expressions based on the driving audio (Figure 5, bottom-left). This is because our multi-modal emotion embedding extracts emotional cues from multiple

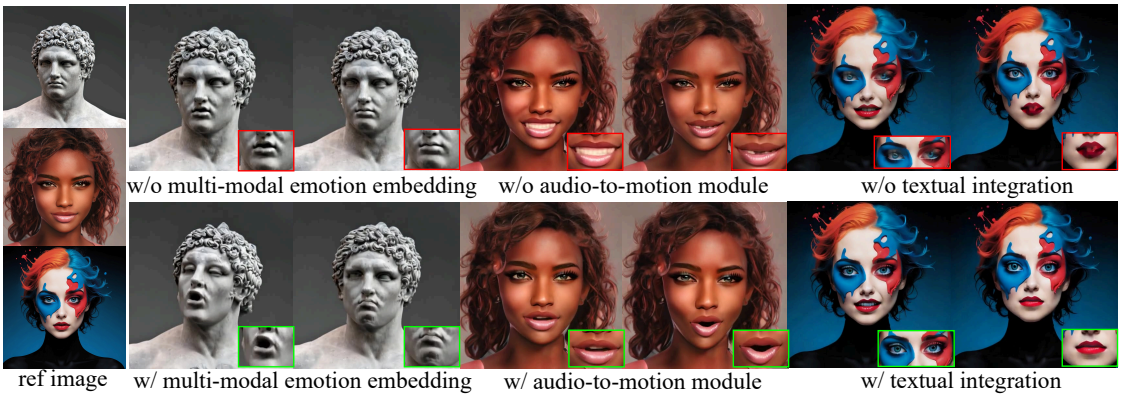


Figure 5. Effects of Multi-modal Emotion Embedding (left), Audio-to-Motion (A2M) (middle) and Textual Integration (right).

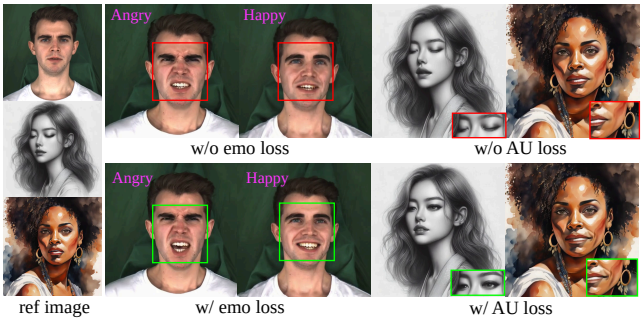


Figure 6. Effects of Emo Loss (left) and AU Loss (right).

Table 5. Effect of adding the multi-modal emotion embedding module and loss functions.

Method	EFID ↓	F1 ↑	CCC_V ↑	CCC_A ↑
w/o emo embedding	1.43	0.58	0.48	0.50
w/o emo loss	1.41	0.63	0.51	0.51
w/o AU loss	1.38	0.63	0.53	0.48
Ours	**1.36**	**0.67**	**0.57**	**0.54**

perspectives, including sentiment analysis, speech emotion recognition, and valence-arousal of the audio.

Audio-to-Motion (A2M) Module. When we omit the A2M module and train our model using only the original video frames, the generated video fails to provide correct synchronization and lacks the necessary subject movements (Figure 5, top-middle). By adding the motion frames driven by audio and training our model on these audio-aware frames, we ensure accurate lip movements. Based on FVD and Sync scores (Table 3), our approach provides better video quality while maintaining the accurate lip synchronization.

LLM based Textual Integration. We evaluate our model without integrating the textual input, relying mainly on appearance cues from the reference image. Without additional semantic guidance from text, the visual quality of the generated video degrades, although lip sync remains unaffected. The model generates inconsistent facial attributes, such as noticeable artifacts in makeup near the eyes and distortions in lip shape (Figure 5, top-right). Table 4 shows the quantitative results with and without textual integration.

Emo Loss. Although the multi-modal emotion embedding module enables the model to generate emotional expressions, we found that training the model without emo loss results in less expressive and less accurate facial emotion (Figure 6, top-left). By including the emo loss, which aligns the model's output with valence and arousal cues derived from audio, the generated expressions become more expressive and realistic (Figure 6, bottom-left). The results in Table 5 validate the effectiveness of the emo loss in our model.

AU Loss. When we exclude the AU loss, the generated faces exhibit limited facial actions; for example, lip region appears flat and expressionless, and eyes remain closed despite the emotional context (Figure 6, top-right). With AU loss, the model is able to generate fine-grained facial movements, such as the appearance of expression lines near the mouth and properly opened eyes, even when the reference image shows closed eyes (Figure 6, bottom-right). The results in Table 5 show that AU loss enhances the realism and expressiveness of the generated videos.

7. Conclusion, Limitations and Future Work

We propose a novel framework that effectively integrates multi-modal emotional nuances with audio-driven motion modules to generate high-quality, lip-synchronized talking face video. By conditioning the model on visual and textual info, we provide better visual details. Comprehensive experiments, ablation studies & user evaluations demonstrate our model outperforms SOTA. Results show that SynchroRaMa is an effective tool for creating high-quality, emotionally rich, and lip-synchronized talking face videos.

While our proposed approach achieves promising results, it has some limitations which we will address in future work. First, since our model is trained mainly on portrait images, it is currently unable to generate full-body talking videos. Additionally, because the model was trained only on English language data, its performance on other languages needs to be evaluated.

References

[1] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024. 2, 3, 6, 7

[2] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 2, 3, 5

[3] Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with diffusion transformer networks. *arXiv preprint arXiv:2412.00733*, 2024. 6

[4] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 6

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[6] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2663–2671, 2022. 2

[7] Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8498–8507, 2024. 2

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[9] Jochen Hartmann. Emotion english distilroberta-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/, 2022. 4

[10] Tianyu He, Junliang Guo, Runyi Yu, Yuchi Wang, Jialiang Zhu, Kaikai An, Leyi Li, Xu Tan, Chunyu Wang, Han Hu, HsiangTao Wu, Sheng Zhao, and Jiang Bian. Gaia: Zero-shot talking avatar generation, 2024. 2

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[12] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 3

[13] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5325–5329. IEEE, 2024. 5

[14] Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and CV Jawahar. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1428–1436, 2019. 2

[15] Tao Liu, Feilong Chen, Shuai Fan, Chenpeng Du, Qi Chen, Xie Chen, and Kai Yu. Anitalker: animate vivid and diverse talking faces through identity-decoupled facial motion encoding. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6696–6705, 2024. 2

[16] Yunfei Liu, Lijian Lin, Fei Yu, Changyin Zhou, and Yu Li. Moda: Mapping-once audio-driven portrait animation with dual attentions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23020–23029, 2023. 3, 6

[17] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines. *CoRR*, abs/1906.08172, 2019. 6

[18] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 2

[19] Soumik Mukhopadhyay, Saksham Suri, Ravi Teja Gadde, and Abhinav Shrivastava. Diff2lip: Audio conditioned diffusion models for lip-synchronization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5292–5302, 2024. 2

[20] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV 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*, pages 484–492, 2020. 2

[21] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. 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*, page 484–492. ACM, 2020. 6

[22] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. 4

[23] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 4

[24] Yi Ren, Jinglin Liu, and Zhou Zhao. Portaspeech: Portable and high-quality generative text-to-speech. *Advances in Neural Information Processing Systems*, 34:13963–13974, 2021. 3

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image

synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[26] Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1982–1991, 2023. 2

[27] Xusen Sun, Longhao Zhang, Hao Zhu, Peng Zhang, Bang Zhang, Xinya Ji, Kangneng Zhou, Daiheng Gao, Liefeng Bo, and Xun Cao. Vividtalk: One-shot audio-driven talking head generation based on 3d hybrid prior. *arXiv preprint arXiv:2312.01841*, 2023. 2

[28] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pages 244–260. Springer, 2024. 2, 6

[29] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10745–10759, 2023. 4, 5

[30] Cong Wang, Kuan Tian, Jun Zhang, Yonghang Guan, Feng Luo, Fei Shen, Zhiwei Jiang, Qing Gu, Xiao Han, and Wei Yang. V-express: Conditional dropout for progressive training of portrait video generation. *arXiv preprint arXiv:2406.02511*, 2024. 6, 7

[31] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change 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, 7

[32] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[33] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024. 2, 3, 6, 7

[34] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 657–666, 2022. 6

[35] You Xie, Hongyi Xu, Guoxian Song, Chao Wang, Yichun Shi, and Linjie Luo. X-portrait: Expressive portrait animation with hierarchical motion attention. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2

[36] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024. 2, 3, 6, 7

[37] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *Advances in Neural Information Processing Systems*, 37:660–684, 2024. 2

[38] Zhenhui Ye, Jinzheng He, Ziyue Jiang, Rongjie Huang, Jiawei Huang, Jinglin Liu, Yi Ren, Xiang Yin, Zejun Ma, and Zhou Zhao. Geneface++: Generalized and stable real-time audio-driven 3d talking face generation. *arXiv preprint arXiv:2305.00787*, 2023. 3

[39] Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, Jinzheng He, and Zhou Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. *arXiv preprint arXiv:2301.13430*, 2023. 3

[40] Phyo Thet Yee, Sudeepta Mishra, and Abhinav Dhall. Clipswap: Towards high fidelity face swapping via attributes and clip-informed loss. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10. IEEE, 2024. 3

[41] Phyo Thet Yee, Sudeepta Mishra, and Abhinav Dhall. Clipswap++: Improved identity and attributes aware face swapping. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2025. 3

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver 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

[43] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie 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. 6