SyncAnimation: A Real-Time End-to-End Framework for Audio-Driven Human Pose and Talking Head Animation

Yujian Liu[1,2], **Shidang Xu**[2], **Jing Guo**[1,3], **Dingbin Wang**[1,4], **Zairan Wang**[1], **Xianfeng Tan**[1] and **Xiaoli Liu**[1,*]
[1]AiShiWeiLai AI Research, Beijing, China
[2]South China University of Technology, Guangzhou, China
[3]Beijing Institute of Technology, Beijing, China
[4]Beijing University of Posts and Telecommunications, Beijing, China

Abstract

Generating talking avatar driven by audio remains a significant challenge. Existing methods typically require high computational costs and often lack sufficient facial detail and realism, making them unsuitable for applications that demand high real-time performance and visual quality. Additionally, while some methods can synchronize lip movement, they still face issues with consistency between facial expressions and upper body movement, particularly during silent periods. In this paper, we introduce SyncAnimation, the first NeRF-based method that achieves audio-driven, stable, and real-time generation of speaking avatar by combining generalized audio-to-pose matching and audio-to-expression synchronization. By integrating AudioPose Syncer and AudioEmotion Syncer, SyncAnimation achieves high-precision poses and expression generation, progressively producing audio-synchronized upper body, head, and lip shapes. Furthermore, the High-Synchronization Human Renderer ensures seamless integration of the head and upper body, and achieves audio-sync lip. The project page can be found at https://syncanimation.github.io/

1 Introduction

In recent years, audio-visual synthesis techniques have garnered significant attention, with audio-driven realistic avatar generation emerging as a key research focus. Over the past few years, many researchers have employed GAN- or SD-based deep generative models to tackle this task [Prajwal *et al.*, 2020; Zhang *et al.*, 2023b; Zhong *et al.*, 2023; Zhang *et al.*, 2023a; Xu *et al.*, 2024; Wang *et al.*, 2024; Chen *et al.*, 2024; Xie *et al.*, 2024; Tan *et al.*, 2025]. Among them, SD-based models, leveraging model parameters and large-scale datasets, can generate fully animated avatars from a single reference image. However, their reliance on large-scale and diverse individual datasets, coupled with overwhelming computational and time costs, limits their applicability in real-time scenarios such as live streaming or video conferencing, where high fidelity and real-time rendering are essential [Guo *et al.*, 2024; Hu, 2024].

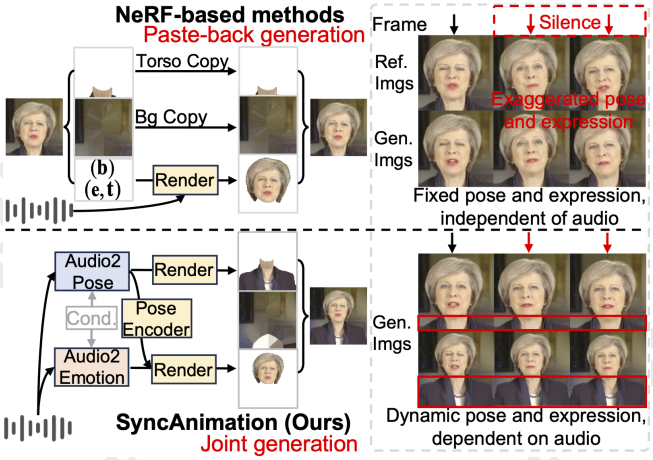

Figure 1: SyncAnimation is the first NeRF-based generative method that uses audio to create realistic facial and upper-body movement.

Recently, Neural Radiance Fields (NeRF) have been applied to audio-driven talking avatars [Mildenhall *et al.*, 2021; Guo *et al.*, 2021; Li *et al.*, 2023; Ye *et al.*, 2023b; Kim *et al.*, 2024; Peng *et al.*, 2024; Ye *et al.*, 2023a]. By associating NeRF with audio either end-to-end or through intermediate representations, these methods enable the reconstruction of personalized talking avatars with impressive synthesis quality and inference speed. However, existing approaches, such as ER-NeRF [Li *et al.*, 2023] and Synctalk [Peng *et al.*, 2024], focus primarily on achieving precise synchronization between lip movement and audio, given strong correlation. Nevertheless, they have yet to tackle the mismatch between audio and head poses, as well as the challenging association between audio and facial expressions [Zhang *et al.*, 2023a], ultimately reducing the realism of the generated avatar.

In this paper, we highlight the importance of generating realistic talking avatars driven by audio, focusing on identity consistency and detail preservation, while facing three critical challenges that require further attention, shown in Fig. 1: (1). Pose inconsistency with audio: Generate identical and fixed poses across different inferred audios (derived from original video frames), possibly even exhibiting exaggerated head

*Corresponding author.

movement in silent segments. (2). Expression inconsistency with audio: Insufficient attention is given to facial animation elements beyond lip-syncing, such as eyebrow movement and blinking, which are crucial to conveying natural expressions and emotional depth, resulting in unnatural and stiff animations. (3). Loss of paste-back ability in audio-driven pose method: Only the head is generated, and the edges of the body are generated with changes, causing the torso to be displaced and therefore cannot be attached back to the origin torso. As mentioned above, audio-driven pose generation methods cannot achieve upper-body generation, and the lack of audio-driven expression generation results in inconsistent expressions. Thus, a fully generative NeRF-based approach with strong audio correlation is essential for achieving realistic talking avatars.

To address these critical challenges, we propose SyncAnimation, a NeRF-based framework focused on audio-driven rendering of upper body and head . This framework integrates three core modules: the AudioPose Syncer and AudioEmotion Syncer, which enable stable, precise, and controllable mappings from audio to head poses and facial expressions, and the High-Synchronization Human Renderer, which ensures seamless integration of head motion and upper-body movement without post-processing. Together, these modules form a unified solution for generating dynamic, expressive, and highly audio-synchronized avatars. In summary, the main contributions of our work are as follows:

- We propose SyncAnimation, a rendering framework for audio-driven expression and upper body generation. This framework generates an avatar that is highly consistent with the audio and displays a diversity of actions, while supporting both one-shot and zero-shot inference.
- We introduce the Audio2Pose and Audio2Emotion modules to support end-to-end efficient training, enabling high-precision poses and expression generation, and progressively generating audio-sync upper body, head, and lip shapes.
- Our algorithm achieves 41 FPS inference on an NVIDIA RTX 4090 GPU, and to our knowledge, this is the first real-time audio-driven avatar method capable of generating audio-sync upper body movement and head motion.
- Extensive experimental results show that SyncAnimation successfully generates realistic avatars with the same scale as the original video and significantly outperforms existing state-of-the-art methods in both quantitative and qualitative evaluations.

2 Related Work

2.1 Paste-Back Generation

GAN- and NeRF-based methods are representative of paste-back generation. Among them, GAN-based talking head synthesis has primarily focused on generating video streams for the lip region, creating new visual effects for talking head avatar[Cheng *et al.*, 2022; Zhang *et al.*, 2023b; Zhong *et al.*, 2023; Tan *et al.*, 2025]. For example, Wav2Lip [Prajwal *et al.*, 2020] introduced a powerful lip-sync discriminator to supervise lip movement and penalize mismatched mouth shapes. IP-LAP [Zhong *et al.*, 2023] proposed an audio-to-landmark generator and a landmark-to-video model, using prior landmark and appearance information to reconstruct lips from a reference image. Recently, EdTalk presented an effective disentanglement framework, using orthogonal bases stored in a dedicated library to represent each spatial component for efficient audio-driven synthesis. With the rise of NeRF, earlier works [Guo *et al.*, 2021; Ye *et al.*, 2023b; Li *et al.*, 2023; Peng *et al.*, 2024] have integrated NeRF into the task of synthesizing talking heads, using audio as the driving signal. For instance, AD-NeRF [Guo *et al.*, 2021] was the first to render both the torso and head but suffered from poor generalization, and the synthesized lip movement sometimes appeared unnatural. ER-NeRF [Li *et al.*, 2023] innovatively introduced tri-plane hash encoders and a region attention module, advocating a fast and precise lip-sync rendering approach. Geneface [Ye *et al.*, 2023b] and SyncTalk [Peng *et al.*, 2024] generated a generalized representation based on extensive 2D audiovisual datasets, ensuring synchronized lip movement across different audios. The GAN-based methods map audio to lip-sync but paste other parts, causing blurry lips (Fig.5). In contrast, NeRF-based methods perform full-face synthesis but fail to synchronize audio with facial expressions, head poses, and upper-body movement.

2.2 Joint Generation

Leveraging the fundamental principles of text-to-image diffusion models, recent advances in video generation based on diffusion techniques have shown promising results [Xu *et al.*, 2024; Wang *et al.*, 2024; Chen *et al.*, 2024]. V-Express [Wang *et al.*, 2024] effectively links lip movement, facial expressions, and head poses through progressive training and conditional dropout operations, enabling precise control using audio. Hallo [Xu *et al.*, 2024] adopts a hierarchical audio-driven visual synthesis approach, achieving lip synchronization, expression diversity, and pose variation control. EchoMimic [Chen *et al.*, 2024] employs a novel training strategy that incorporates both audios and facial landmarks for avatar synthesis. The SD-based methods generate audio-driven talking avatars with poses and facial animations. However, large-scale training often results in poor resemblance to the original individual and comes with high computational costs, making real-time applications like live streaming and video conferencing challenging. For instance, generating a one-minute video can take up to half an hour. Additionally, these methods struggle with audio-motion mismatches when handling out-of-domain audio, further limiting their suitability for real-time use.

3 Method

In this section, we propose SyncAnimation, a generative, audio-driven model for creating avatars with dynamic head motion and upper-body movement in Fig. 2. The framework includes three main components: (1). the AudioPose Syncer for accurate audio-to-head pose mapping in Sec. 3.1, (2). the AudioEmotion Syncer for controllable, audio-driven facial expressions in Sec. 3.2, and (3). the High-Synchronization Human Renderer for seamless upper-body generation without head pasting in Sec. 3.3.

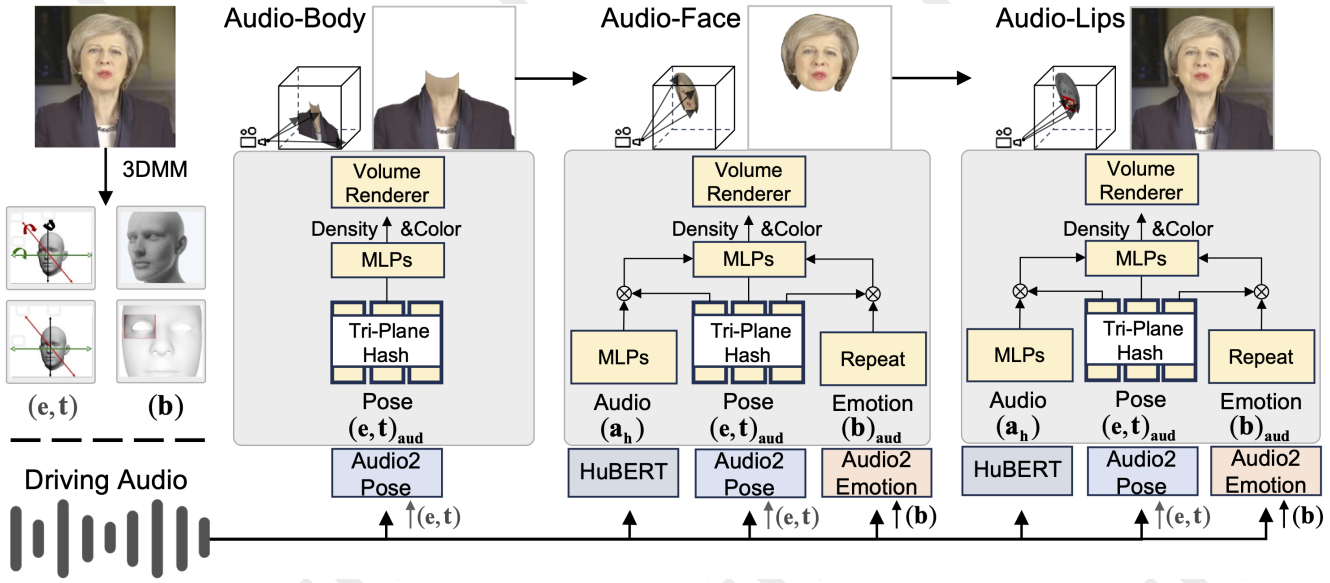

Figure 2: SyncAnimation Framework: Given an image and audio, the preprocessing extracts 3DMM parameters for Audio2Pose and Audio2Emotion as references (or noise). It then generates the upper body, head, and lip refinement, ensuring pose consistency and facial expression alignment with the audio.

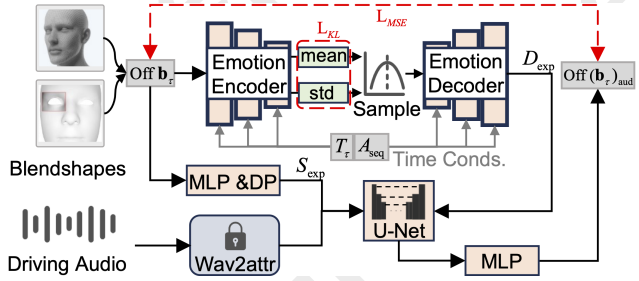

Figure 3: Audio2Pose reconstructs stable head pose offsets using audio and monocular input. It employs the pre-trained Wav2attr audio encoder for person-specific audio encoding, integrates a gaussian-based VAE for diversity, and uses a stability model with high dropout rate for improved pose reconstruction.

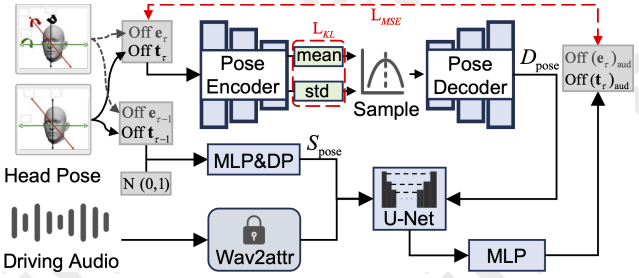

Figure 4: Audio2Emotion reconstructs 3DMM expression offsets from audio and monocular input. To address weak expression-audio correlation and blinking periodicity, we replace the diversity template with a conditional VAE guided by time $\mathbf{T}_\tau$ and context-dependent audio features $\mathbf{A}_{\text{seq}}$.

3.1 AudioPose Syncer

Audio-aware Poses Generation. Learning a model to control head motion from audio is challenging due to two main issues: (1). The rotation matrix must be orthogonal with a determinant of 1, but direct prediction can lead to non-orthogonal matrices, causing errors in rotation and pose reconstruction. (2). There is inherent ambiguity between audio and motion, leading to jitter or frame skipping in the generated avatar. We propose AudioPose to address these issues, as shown in Fig.3. Considering the difficulty in predicting orthogonal rotation matrices with a high-dimensional compact representation, we transform the task into predicting three euler angles (roll α, pitch β, and yaw γ). Given rotation matrix $\mathbf{R}$, the conversion formula is:

$$\alpha = \text{atan2}(\mathbf{R}_{21}, \mathbf{R}_{11}), \quad \beta = \text{atan2}(\mathbf{R}_{32}, \mathbf{R}_{33}),$$
$$\gamma = \text{atan2}(-\mathbf{R}_{31}, \sqrt{\mathbf{R}_{32}^2 + \mathbf{R}_{33}^2}) \tag{1}$$

$\mathbf{R}_{\text{ij}}$ denotes the element in the i-th row and j-th column of the rotation matrix $\mathbf{R}$. The euler angles $\mathbf{e} = (\alpha, \beta, \gamma)$ and the translation $\mathbf{t} = (x, y, z)$ jointly define the head poses, controlling its orientation and position.

The non-linear transformation of the rotation matrix introduces new challenges, such as gimbal lock ($\beta \approx \pm 90°$) causing non-invertibility, and pose variations influenced by euler angles and translations within narrow ranges. In real-world videos, head poses are confined to small ranges, not extending to extreme values like $\pm 90°$ or $\pm 180°$. This insight is incorporated into training to address these challenges. Specially, the prediction range is restricted from $(\mathbf{e}, \mathbf{t}) \in (-\infty, +\infty)$ to $(\Delta\mathbf{e}, \Delta\mathbf{t}) \in (\bar{\mathbf{e}} \pm \Delta, \bar{\mathbf{t}} \pm \Delta)$, where $\bar{\mathbf{e}}$ and $\bar{\mathbf{t}}$ are the average poses, and Δ defines the deviation. To ensure consistency, outputs are converted into a normalized distribution $(\bar{\Delta\mathbf{e}}, \sigma_{\Delta\mathbf{e}})$, standardizing the deviation range as follows:

$$\textbf{Off}\,(\mathbf{e}) = \frac{\Delta\mathbf{e} - \bar{\Delta\mathbf{e}}}{\sigma_{\Delta\mathbf{e}}}, \quad \textbf{Off}\,(\mathbf{t}) = \frac{\Delta\mathbf{t} - \bar{\Delta\mathbf{t}}}{\sigma_{\Delta\mathbf{t}}} \tag{2}$$

To address the ambiguity between audio and pose in issue 2, we introduce two conditional vectors $\mathbf{S}_{\text{pose}}$ and $\mathbf{D}_{\text{pose}}$. The overall Audio2Pose prediction pipeline is as follows:

$$\textbf{Off}\,(\mathbf{e})_{\text{aud}}, \textbf{Off}\,(\mathbf{t})_{\text{aud}} = F(g(\mathbf{a}), \mathbf{D}_{\text{pose}}, \mathbf{S}_{\text{pose}}) \quad (3)$$

Where $F(\cdot)$ represents the pose generation model. Both $\mathbf{D}_{\text{pose}}$ and $\mathbf{S}_{\text{pose}}$ are conditional vectors, with $\mathbf{S}_{\text{pose}}$ enhancing pose stability and $\mathbf{D}_{\text{pose}}$ guiding diversity. For $g(\mathbf{a})$, we use FaceXHuBERT [Haque and Yumak, 2023] as the audio encoder, as it is more suited than pre-trained models like DeepSpeech and wav2vec [Hannun *et al.*, 2014; Baevski *et al.*, 2020], which lack person-specific information.

Audio-only poses prediction $\left(\mathbf{a} \rightarrow \textbf{Off}\,(\mathbf{e})_{\text{aud}}, \textbf{Off}\,(\mathbf{t})_{\text{aud}}\right)$ suffers from ambiguity. To address this, we introduce an conditional vector $\mathbf{S}_{\text{pose}}$ to enhance pose stability by providing reference poses, reducing the solution space from one-to-many to many-to-one. During training, the regression loss updates $\mathbf{S}_{\text{pose}}$, which is adaptively adjusted to resolve ambiguities. Specifically, at the τ-th frame, gaussian noise is added to the poses of the $(\tau-1)$-th frame, which is then encoded through Multilayer Perceptrons (MLPs) to generate $\mathbf{S}_{\text{pose}}$.

$$\mathbf{S}_{\text{pose}} = f_{\text{MLPs}}\big([\textbf{Off}\,(\mathbf{e}_{\tau-1}), \textbf{Off}\,(\mathbf{t}_{\tau-1})] + \mathcal{N}(\mu_S, \delta_S)\big) \quad (4)$$

where $\textbf{Off}\,(\mathbf{e}_{\tau-1})$ and $\textbf{Off}\,(\mathbf{t}_{\tau-1})$ represent the pose information of the $(\tau-1)$-th frame, and $\mathcal{N}(\mu_S, \delta_S)$ denotes gaussian noise with mean μ_S and standard deviation δ_S.

To generate stable yet diverse poses, We further introduce a diversity-guided conditional vector $\mathbf{D}_{\text{pose}}$, which injects uncertainty and provides diverse pose templates via Gaussian sampling. VAE backpropagation improves convergence and accuracy over single audio regression (Exp. 4.5). To normalize the $\mathbf{D}_{\text{pose}}$ vector space, we apply the KL-divergence loss:

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}\big(\mathcal{N}(\boldsymbol{\mu_D}, \boldsymbol{\delta_D^2}) || \mathcal{N}(0, \mathbf{I})\big) \quad (5)$$

where $\mathcal{N}(\boldsymbol{\mu_D}, \boldsymbol{\delta_D^2}) \in \mathcal{R}^D$ is the latent distribution, and $\mathcal{N}(0, \mathbf{I})$ is the standard normal distribution. For the poses generation model $F(\cdot)$, we use the U-Net architecture [Qian *et al.*, 2021; Wang *et al.*, 2024], which extracts audio features and incorporates $\mathbf{D}_{\text{pose}}$ and $\mathbf{S}_{\text{pose}}$. The U-Net generates a shared feature representation, which is then processed by MLPs to predict rotation $\textbf{Off}\,(\mathbf{e})_{\text{aud}}$ and translation $\textbf{Off}\,(\mathbf{t})_{\text{aud}}$. Reconstruction loss $\mathcal{L}_{\text{reg}}$ is applied to align the generated poses with the ground truth.

$$\mathcal{L}_{\text{reg}} = \|\big(\textbf{Off}\,(\mathbf{e}), \textbf{Off}\,(\mathbf{t})\big) - \big(\textbf{Off}\,(\mathbf{e})_{\text{aud}}, \textbf{Off}\,(\mathbf{t})_{\text{aud}}\big)\|_1 \quad (6)$$

Therefore, the final poses generation loss is:

$$\mathcal{L}_{\text{pose}} = \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \quad (7)$$

where λ_{KL} and λ_{reg} are the weights for the VAE and regression loss terms, respectively.

Audio-guided Upper-body Generation. We generally follow the rendering process in preceding work [Guo *et al.*, 2021; Li *et al.*, 2023]. Originated from the representation of neural radiance field, the implicit function is defined as $\mathcal{F}_{\text{NeRF}} : (\mathrm{x}, d) \rightarrow (c, \sigma)$, where $\mathrm{x} = (x, y, z)$ refers to 3D spatial location and d is the viewing direction. The output c and σ determine color value and corresponding density.

To render high-fidelity and realistic scene efficiently, we employ a 2D-multiresolution hash encoder introduced by [Li *et al.*, 2023]. A specific encoder designed for position $\mathrm{x} = (x, y, z)$ projected on plane AB is defined as $\mathcal{H}^{AB} : (a, b) \rightarrow h_{ab}^{\text{AB}}$, where a and b are projected coordinates, while $h_{ab}^{\text{AB}} \in \mathbf{R}^{LD}$ represents L levels of features with D dimensions.

To ensure dynamic upper-body generation and consistency between head motion and upper body movement, the audio-predicted $\textbf{Off}\,(\mathbf{e})_{\text{aud}}$ and $\textbf{Off}\,(\mathbf{t})_{\text{aud}}$ are recovered to space representations $\mathbf{e}_{\text{aud}}$ and $\mathbf{t}_{\text{aud}}$ via inverse normalization. Subsequently, three trainable coordinates $\tilde{X}$ are deformed as rendering conditions. The implicit function is defined as:

$$F_{\text{upper-body}} : (\mathrm{x}, \mathbf{e}_{\text{aud}}, \mathbf{t}_{\text{aud}}, \tilde{X}, \mathcal{H}^t) \rightarrow (c, \sigma), \quad (8)$$

where $\mathcal{H}^t$ denote the hash encoder. During training, we optimize the upper-body model by minimizing the error between $\hat{\mathcal{C}}(r)$ and genuine pixel color $\mathcal{C}(r)$ through:

$$\mathcal{L}_{\text{upper-body}} = ||\mathcal{C}(r) - \hat{\mathcal{C}}(r)||_2^2. \quad (9)$$

3.2 AudioEmotion Syncer

Lifelike Expression Prediction. Audio-driven research often focuses on lip-sync, neglecting other facial expressions with weaker audio correlation [Tian *et al.*, 2025], leading to unnatural expressions. We use Arkit face blendshapes, semantically rich 3D coefficients, to model the upper face region $\mathbf{b}$ [Peng *et al.*, 2023]. However, for the eyes, $\mathbf{b}$ reaches extreme values (0 or 1) during full eye closure or opening. To enable full eye closure, we predict the offsets $\textbf{Off}\,(\mathbf{b})$ relative to the average $\bar{\mathbf{b}}$ and incorporate stability constraints $\mathbf{S}_{\text{exp}}$ and diversity guidance $\mathbf{D}_{\text{exp}}$ (Fig.4).

$$\textbf{Off}\,(\mathbf{b})_{\text{aud}} = F(g(\mathbf{a}), \mathbf{D}_{\text{exp}}, \mathbf{S}_{\text{exp}}) \quad (10)$$

Blinking relies on temporal patterns, making it hard to model and often causing unnatural behavior like missing blinks. To address this, we enhance the VAE within the diversity-guided $\mathbf{D}_{\text{exp}}$ to capture temporal dependencies by incorporating time features $\mathbf{T}_\tau$ and sequential audio $\mathbf{A}_{\text{seq}}$, enabling the CVAE to generate $\mathbf{D}_{\text{exp}}$ with strong temporal correlations.

$$\mathbf{D}_{\text{exp}} = f_{\text{CVAE}}\big(\textbf{Off}(\mathbf{b}) \mid T_\tau, \mathbf{A}_{\text{seq}}\big) \quad (11)$$

where $\mathbf{T}_\tau$ is encoded via sinusoidal functions and processed by an MLP. $\mathbf{A}_{\text{seq}}$ is derived by integrating audio data from n neighboring frames $\{\mathbf{a}_{\tau-n}, \ldots, \mathbf{a}_{\tau+n}\}$, and processed through stacked convolutional layers to generate temporally contextualized representations that ensure dynamic expression continuity. Given the weak audio-expression correlation, we enhance the expression reference by modifying the input to the stability vector $\mathbf{S}_{\text{exp}}$. Specifically, the previous expression $\mathbf{b}_{\tau-1}$ and gaussian noise $\mathcal{N}(\mu_S, \delta_S)$ are replaced by the current frame's expression $\mathbf{b}_\tau$.

For the audio-driven expression model, we optimize L_{exp} using the following loss function:

$$\mathcal{L}_{\text{exp}} = \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \quad (12)$$

where λ_{KL} and λ_{reg} are weights applied to the CAVE and regression loss terms.

Method		Image Quality				Lip Sync			Head Motion	
		PSNR↑	LPIPS↓	SSIM↑	FID↓	LMD↓	AUE↓	SyncScore↑	Diversity↑	EAR↓
GAN	Wav2Lip	18.8071	0.2523	0.6571	40.3604	5.2161	4.5499	**9.2332**	0.0782	0.0535
	DINet	18.6620	0.2547	0.6515	34.0215	5.2784	4.4457	7.3138	0.0560	0.0568
	IP-LAP	18.7763	0.2438	0.6531	34.8322	5.4905	4.7474	3.4331	0.0444	0.0545
	EDTalk	18.7482	0.2896	0.6670	53.7573	5.0925	4.8511	7.3092	0.1046	0.0441
Nerf	ER-NeRF	18.8610	0.2323	0.6799	37.5604	3.7192	4.8899	6.1909	0.0935	0.0439
	SyncTalk	18.8035	0.2287	0.6766	32.9779	3.6106	3.6198	7.0865	0.0822	0.0512
	GeneFace++	19.0344	0.2443	0.6819	40.5045	5.3823	4.0530	7.1118	0.0726	0.0504
SD	Hallo	17.9117	0.2678	0.6305	35.8346	6.0491	4.3952	7.2458	0.2166	0.0471
	V-Express	17.3918	0.2842	0.6121	46.6975	6.9958	5.4740	7.4739	0.1039	0.0490
	EchoMimic	14.4399	0.4524	0.4562	59.3261	8.0643	5.1434	6.0574	0.1864	0.0468
SyncAnimation-One		21.2323	0.1543	0.7343	**20.0567**	3.2215	**3.5485**	7.1389	0.2570	**0.0357**
SyncAnimation-Zero		**21.4006**	**0.1532**	**0.7411**	21.7353	**3.1878**	3.5515	7.1499	**0.2652**	0.0386

Table 1: Quantitative comparisons with state-of-the-art methods. "SyncAnimation-One" refers to one-shot inference, while "SyncAnimation-Zero" indicates zero-shot inference using noise of the same dimension. We achieve state-of-the-art performance on most metrics.

Dynamic Head Rendering. We generate realistic head poses using a technique similar to Sec.3.1. The audio-predicted **Off** $\mathbf{e}_{\text{aud}}$, **Off** $\mathbf{t}_{\text{aud}}$, and **Off** $\mathbf{b}_{\text{aud}}$ are converted to $\mathbf{e}_{\text{aud}}$, $\mathbf{t}_{\text{aud}}$, and $\mathbf{b}_{\text{aud}}$ in the original space via inverse normalization. Due to the complexity of head generation, we use three 2D hash encoders, each operating on a specific plane. The final geometric feature is obtained by concatenating the outputs from all three planes. Predicted coefficients control expressions, and audio features enhance head motion prediction. The implicit function is then defined as:

$$F_{\text{head}} : (\mathrm{x}, \mathbf{e}_{\text{aud}}, \mathbf{t}_{\text{aud}}, \mathbf{b}_{\text{aud}}, a_{\text{h}}, \mathcal{H}^3) \rightarrow (c, \sigma), \qquad (13)$$

Here, a_{h} is the audio representation extracted by Hubert [Hsu *et al.*, 2021], distinct from the one used for predicting poses and expressions due to different task requirements. The corresponding loss $\mathcal{L}_{\text{head}}$ aligns with $\mathcal{L}_{\text{upper-body}}$.

3.3 High-Synchronization Human Renderer

Facial-aware attention. To better utilize audio and expression information in different facial regions, we apply a channel-wise attention mechanism [Li *et al.*, 2023]. From the hash encoders' output, we obtain attention weights for audio and expression using two MLPs: Attn_{aud} and Attn_{exp}.

$$v_{\text{aud}} = \text{Attn}_{\text{aud}}(\mathcal{H}^3(x)),\ v_{\text{exp}} = \text{Attn}_{\text{exp}}(\mathcal{H}^3(x)) \qquad (14)$$

By applying hadamard product we attain region-aware features $a_{h,x} = a_{h,x} \odot v_{\text{aud}}$ and $\mathbf{b}_{\text{out},x} = \mathbf{b}_{\text{out}} \odot v_{\text{exp}}$. Such operation makes sure model can explore useful information in a disentangled way and thus increase rendering quality.

Fine Lip Optimization. Though audio feature benefits holistic head rendering, we seek to manually augment attention weight on lip region due to more relevance. By utilizing mask technique [Peng *et al.*, 2024], we lower the attention weight out of lip area. Meanwhile, we use LPIPS loss focused on lip zone to gain finer result. The process is expressed as:

$$\mathcal{L}_{\text{lip}} = \mathcal{L}_{\text{head}} + \lambda \text{LPIPS}(\mathcal{P}, \hat{\mathcal{P}}). \qquad (15)$$

4 Experiments

4.1 Implementations

Dataset. For fair comparison, the experimental dataset was obtained from publicly available video collections in [Guo *et al.*, 2021; Li *et al.*, 2023; Peng *et al.*, 2024] and HDTF [Zhang *et al.*, 2021]. We collected well-edited video sequences in English, French, and Korean, with an average of 6665 frames per video at 25 FPS. Each raw video was standardized to 512×512, with the a center portrait.

Quantitative Evaluation Metrics We evaluate our method using widely adopted metrics. For image quality, we employ full-reference metrics, including Peak Signal-to-Noise Ratio (PSNR) [Hore and Ziou, 2010], Learned Perceptual Image Patch Similarity (LPIPS) [Zhang *et al.*, 2018], Structural Similarity Index Measure (SSIM) [Wang *et al.*, 2004], and Frechet Inception Distance (FID) [Heusel *et al.*, 2017]. In terms of lip and face synchronization, we utilize landmark distance (LMD) to measure the synchronicity of facial movement [Chen *et al.*, 2018], and Action Unit Error (AUE) to evaluate the accuracy of facial expressions [Baltrušaitis *et al.*, 2015]. Furthermore, we introduce Lip Sync Error Confidence (LSE-C), consistent with Wav2Lip, to evaluate the synchronization between lip movement and audio [Prajwal *et al.*, 2020].For the diversity of head motion generated, a standard deviation of the embeddings of head motion features is extracted from the generated frames using Hopenet [Ruiz *et al.*, 2018]. For eye blink detection, The Eye Aspect Ratio (EAR), calculated based on the positions of facial landmarks around the eyes, is utilized to evaluate the naturalness and accuracy of generated eye movement [Soukupova and Cech, 2016].

Comparison Baselines. We compare SyncAnimation with two SOTA methods for lip synchronization and motion generation. For lip synchronization, we include GAN-based methods like Wav2Lip [Prajwal *et al.*, 2020], DInet [Zhang *et al.*, 2023b], and IP-LAP [Zhong *et al.*, 2023], as well as NeRF-based methods such as SyncTalk [Peng *et al.*, 2024], AD-NeRF [Guo *et al.*, 2021], ER-NeRF [Li *et al.*, 2023], and GeneFace++ [Ye *et al.*, 2023a]. For motion generation, we compare SyncAnimation with three large-scale stable diffusion models: V-Express [Wang *et al.*, 2024], Hallo [Xu *et al.*, 2024], and EchoMimic [Chen *et al.*, 2024].

Implementation Details. We train SyncAnimation with separate steps for upper-body generation, head rendering, and lip-audio synchronization, using 15k, 12k, and 4k steps, respectively. Each iteration samples 256^2 rays and employs a 2D hash encoder with parameters $L = 14$ and $F = 1$.

Figure 5: Visual comparison with outputs of baselines. GAN- and NeRF-based methods generate avatar with fixed poses and expressions. SD-based methods allow expression changes but lack facial detail and pose movement. SyncAnimation uniquely achieves jointly generative, audio-driven realistic expressions and movable poses.

The AdamW optimizer is used, with learning rates set to 0.01 for the hash encoder and 0.001 for the Audio2Pose and Audio2Emotion modules. The total training time is approximately 4 hours on an NVIDIA RTX 4090 GPU.

4.2 Quantitative Evaluation

Compare with baseline. We first compare one-shot SyncAnimation (SyncAnimation-One) with several state-of-the-art methods using the first frame and test audio for driving. As shown in Tab.1, SyncAnimation-One outperforms GAN-, NeRF-, and SD-based approaches. (1). In terms of image quality, SyncAnimation-One achieves superior results in PSNR, LMD, and FID, indicating better detail preservation. (2). For synchronization, SyncAnimation-One surpasses other methods on LMD and AUE, demonstrating excellent audio-lip synchronization. Although SyncScore lags behind GAN-based methods due to SyncNet training loss, it still outperforms NeRF-based methods. (3). SyncAnimation-One significantly outperforms GAN-, NeRF-, and even SD-based methods and achieves the lowest EAR error.

Our framework supports both one-shot and zero-shot inference. In zero-shot mode (SyncAnimation-Zero), reference poses and blendshapes are replaced with gaussian noise. As shown in Tab.1, SyncAnimation-One (using the first frame as reference) and SyncAnimation-Zero (using gaussian noise) perform similarly across all metrics and outperform other methods, demonstrating that SyncAnimation operates independently of reference inputs for identity or information, relying solely on audio cues for dynamic rendering.

4.3 Qualitative Evaluation

In the previous section, we quantitatively demonstrated the superiority of SyncAnimation. In this section, we visually assess the quality of generated frames by comparing GAN-, NeRF-, and SD-based methods with SyncAnimation, as illustrated in Fig.5. Unlike other methods that render the upper body but fail to synchronize movement with audio, SyncAnimation excels by producing audio-synced upper-body movement. Additionally, GAN- and NeRF-based methods struggle with upper facial expressions during one-shot inference. Although SD-based methods generate diverse facial expressions, their reliance on large-scale training often fails to capture fine details like eyes and lips, causing issues such as asymmetry, overexposure, missing teeth, unfocused eyes, and unnatural lip closures. In contrast, SyncAnimation preserves the subject's identity with superior fidelity and resolution, accurately reproducing subtle actions like blinking and eyebrow movement, thanks to the AudioEmotion Syncer. For a more comprehensive comparison, we recommend watching the supplementary video.

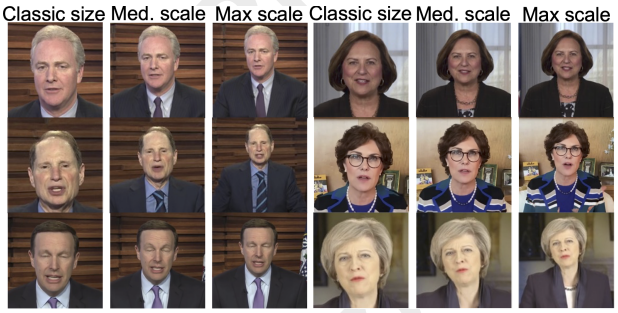

Figure 6: Visual generation results of the proposed method given different upper-body scaling expansion

4.4 Upper-body Scaling Expansion

In real-time applications (e.g., news broadcasting and live teaching), individuals often appear in upper-body. Existing

Method	Image Quality				Lip Sync			Head Motion	
	PSNR	LPIPS	SSIM	FID	LMD	AUE	SyncScore	Diversity	EAR
SyncAnimation	20.1038	0.1860	0.6752	26.8092	3.0746	**2.8812**	**7.8717**	**0.2443**	**0.0405**
SyncAnimation-MedScale	20.2972	0.1685	0.7051	27.4110	2.7120	3.7966	7.2024	0.2158	0.0413
SyncAnimation-MaxScale	**21.5558**	**0.1192**	**0.7761**	**22.3477**	**2.0754**	4.0455	7.3364	0.1632	0.0406

Table 2: Qualitative comparison with varying upper-body scales

methods [Ye *et al.*, 2023a; Peng *et al.*, 2024] often paste the rendered head onto the original torso, resulting in unnatural avatars and limiting applicability. There is a growing demand for direct upper-body rendering that is highly synchronized with audio.

We depart from the conventional approach of restricting upper-body scale and pasting back to the original frame. SyncAnimation directly renders upper-body avatars driven by audio, progressively increasing the upper body proportion in the rendered images. Image quality and audio consistency metrics are shown in Tab.2. As the upper-body scale increases,

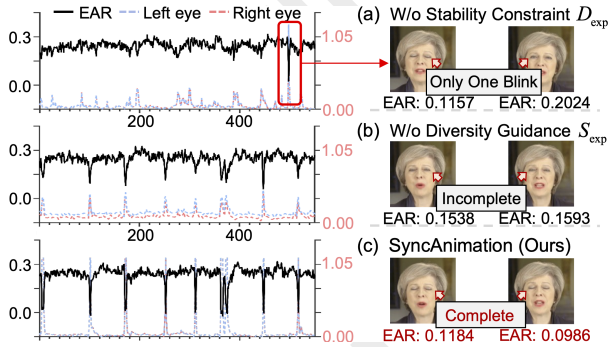

Figure 7: Ablation Study on $\mathbf{D}_{\text{exp}}$ and $\mathbf{S}_{\text{exp}}$ in OOD audio inference. Removing them will result in (a) and (b).

overall image quality improves, but lip-sync consistency and head motion diversity decline. This suggests that while the upper body is easier to render, a higher upper-body proportion improves image quality, but reduces focus on the head. The EAR remains stable due to the periodic time features in $\mathbf{S}_{\text{exp}}$ from Audio2Emotion. As shown in Fig.6, we scale the rendered avatar to match the original video's size, demonstrating that SyncAnimation can generate upper-body avatars with strong audio correlation and natural poses.

4.5 Ablation Study

In this section, we report ablation studies under the joint generation setting to validate the effectiveness of our major contributions from two perspectives. Additionally, to demonstrate the adaptability of the SyncAnimation to Out-of-Domain (OOD) audio, we evaluate the impact of different backbones on rendering structures under external audio conditions. The results are presented in Tab.3 and Fig.7.

Effectiveness of $\mathbf{D}_{\text{pose}}$ and $\mathbf{S}_{\text{pose}}$

The Audio2Pose model is key to generating stable and natural free-form poses. To demonstrate the effectiveness of the diversity-guided conditional vector $\mathbf{D}_{\text{pose}}$ and stability-guided vector $\mathbf{S}_{\text{pose}}$, we evaluate the density metric in static and dynamic audio, as well as the absolute error at final convergence. As shown in Tab. 3, incorporating $\mathbf{D}_{\text{pose}}$ and $\mathbf{S}_{\text{pose}}$ either individually or together consistently produces natural avatar. However, the absence of $\mathbf{S}_{\text{pose}}$ leads to unstable head generation with frequent jittering, resulting in an increased density metric. On the other hand, excluding $\mathbf{D}_{\text{pose}}$ results in larger absolute poses errors, and the diversity in silent segments becomes excessively high, making it difficult to achieve the desired small-scale head motion.

Effectiveness of $\mathbf{D}_{\text{exp}}$ and $\mathbf{S}_{\text{exp}}$

In the Audio-driven talking avatar, we evaluate the impact of the proposed $\mathbf{D}_{\text{exp}}$ and $\mathbf{S}_{\text{exp}}$ on facial expressions, specifically focusing on blinking behavior. The evaluation metric is the EAR (Eye Aspect Ratio) calculated for each frame, along with line plots of blendshapes for the left and right eyes predicted by the Auido2Pose model. As shown in Fig.7, when $\mathbf{D}_{\text{exp}}$ or $\mathbf{S}_{\text{exp}}$ is not incorporated, the rendered image exhibits either a single blink over the 22-second driven audio or incomplete eye closures despite EAR fluctuations. Under the combined effect of $\mathbf{D}_{\text{exp}}$ and diversity-guided $\mathbf{S}_{\text{exp}}$, periodic fluctuations are guided by $\mathbf{D}_{\text{exp}}$, while $\mathbf{S}_{\text{exp}}$ accentuates each peak (as illustrated in the comparison between (b) and (c)), resulting in more natural and rhythmic blinking motion.

Method	**L1-error ↓ (latest iter)**	**Diversity ↑ (Normal)**	**Diversity↓ (Silence)**
w/o $\mathbf{D}_{\text{pose}}$,$\mathbf{S}_{\text{pose}}$	1.4910	None	None
w/o $\mathbf{D}_{\text{pose}}$	0.1677	0.3333	0.2142
w/o $\mathbf{S}_{\text{pose}}$	0.009	0.3067	0.1584
Ours	0.009	0.3117	0.1209

Table 3: Ablation the diversity and L1 error of the proposed $\mathbf{D}_{\text{pose}}$ and $\mathbf{S}_{\text{pose}}$ in OOD audio inference. Each conditional template contribute largely to generate realistic head motion.

5 Conclusion

In this paper, we propose SyncAnimation, which aims to achieve three key objectives for modern talking avatars: generalized audio-to-head poses matching, consistent audio-to-facial expression synchronization, and jointly generative upper-body avatars. Our framework integrates AudioPose Syncer, AudioEmotion Syncer, and High-Synchronization Human Renderer modules. These components enable the generation of stable synchronized poses, upper-body movement, and realistic expressions while preserving subject identity, even from monocular or noisy inputs. Extensive experiments demonstrate that our method achieves the three goals of modern talking avatars and outperforms existing methods.

Acknowledgements

Yujian Liu and Shidang Xu have made equal contributions to SyncAnimation.

References

[Baevski *et al.*, 2020] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[Baltrušaitis *et al.*, 2015] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, volume 6, pages 1–6. IEEE, 2015.

[Chen *et al.*, 2018] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European conference on computer vision (ECCV)*, pages 520–535, 2018.

[Chen *et al.*, 2024] 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.

[Cheng *et al.*, 2022] Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, and Nannan Wang. Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022.

[Guo *et al.*, 2021] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5784–5794, 2021.

[Guo *et al.*, 2024] Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Pengfei Wan, Di Zhang, Yufan Liu, Weiming Hu, Zhengjun Zha, et al. I2v-adapter: A general image-to-video adapter for diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024.

[Hannun *et al.*, 2014] Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Gregory Frederick Diamos, Erich Elsen, Ryan J. Prenger, Sanjeev Satheesh, Shubho Sengupta, Vinay Rao, Adam Coates, and A. Ng. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[Haque and Yumak, 2023] Kazi Injamamul Haque and Zerrin Yumak. Facexhubert: Text-less speech-driven e(x)pressive 3d facial animation synthesis using self-supervised speech representation learning. In *Proceedings of the 25th International Conference on Multimodal Interaction*, ICMI '23, page 282–291. Association for Computing Machinery, 2023.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[Hore and Ziou, 2010] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.

[Hsu *et al.*, 2021] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, page 3451–3460, October 2021.

[Hu, 2024] 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.

[Kim *et al.*, 2024] Gihoon Kim, Kwanggyoon Seo, Sihun Cha, and Junyong Noh. Nerffacespeech: One-shot audio-diven 3d talking head synthesis via generative prior. *arXiv preprint arXiv:2405.05749*, 2024.

[Li *et al.*, 2023] Jiahe Li, Jiawei Zhang, Xiao Bai, Jun Zhou, and Lin Gu. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7568–7578, 2023.

[Mildenhall *et al.*, 2021] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[Peng *et al.*, 2023] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin 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.

[Peng *et al.*, 2024] Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Jun He, Hongyan Liu, and Zhaoxin Fan. Synctalk: The devil is in the synchronization for talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 666–676, 2024.

[Prajwal *et al.*, 2020] 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.

[Qian *et al.*, 2021] Shenhan Qian, Zhi Tu, Yihao Zhi, Wen Liu, and Shenghua Gao. Speech drives templates: Co-speech gesture synthesis with learned templates. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11077–11086, 2021.

[Ruiz *et al.*, 2018] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2074–2083, 2018.

[Soukupova and Cech, 2016] Tereza Soukupova and Jan Cech. Eye blink detection using facial landmarks. In *21st computer vision winter workshop, Rimske Toplice, Slovenia*, volume 2, 2016.

[Tan *et al.*, 2025] Shuai Tan, Bin Ji, Mengxiao Bi, and Ye Pan. Edtalk: Efficient disentanglement for emotional talking head synthesis. In *European Conference on Computer Vision*, pages 398–416. Springer, 2025.

[Tian *et al.*, 2025] 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, 2025.

[Wang *et al.*, 2004] 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.

[Wang *et al.*, 2024] 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.

[Xie *et al.*, 2024] 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.

[Xu *et al.*, 2024] 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.

[Ye *et al.*, 2023a] 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.

[Ye *et al.*, 2023b] 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.

[Zhang *et al.*, 2018] 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.

[Zhang *et al.*, 2021] 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.

[Zhang *et al.*, 2023a] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei 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.

[Zhang *et al.*, 2023b] Zhimeng Zhang, Zhipeng Hu, Wenjin Deng, Changjie Fan, Tangjie Lv, and Yu Ding. Dinet: Deformation inpainting network for realistic face visually dubbing on high resolution video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3543–3551, 2023.

[Zhong *et al.*, 2023] Weizhi Zhong, Chaowei Fang, Yinqi Cai, Pengxu Wei, Gangming Zhao, Liang Lin, and Guanbin Li. Identity-preserving talking face generation with landmark and appearance priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2023.