GoHD: Gaze-oriented and Highly Disentangled Portrait Animation with Rhythmic Poses and Realistic Expressions

Ziqi Zhou[1,2], Weize Quan[1,2*], Hailin Shi[3], Wei Li[4], Lili Wang[5], Dong-Ming Yan[1,2]

[1]MAIS, Institute of Automation, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]NIO
[4]Banma
[5]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University

Abstract

Audio-driven talking head generation necessitates seamless integration of audio and visual data amidst the challenges posed by diverse input portraits and intricate correlations between audio and facial motions. In response, we propose a robust framework GoHD designed to produce highly realistic, expressive, and controllable portrait videos from any reference identity with any motion. GoHD innovates with three key modules: Firstly, an animation module utilizing latent navigation is introduced to improve the generalization ability across unseen input styles. This module achieves high disentanglement of motion and identity, and it also incorporates gaze orientation to rectify unnatural eye movements that were previously overlooked. Secondly, a conformer-structured conditional diffusion model is designed to guarantee head poses that are aware of prosody. Thirdly, to estimate lip-synchronized and realistic expressions from the input audio within limited training data, a two-stage training strategy is devised to decouple frequent and frame-wise lip motion distillation from the generation of other more temporally dependent but less audio-related motions, e.g., blinks and frowns. Extensive experiments validate GoHD's advanced generalization capabilities, demonstrating its effectiveness in generating realistic talking face results on arbitrary subjects.

Code — https://github.com/Jia1018/GoHD

1 Introduction

Audio-driven portrait animation, widely applied in social media and mixed reality contexts like avatar creation and teleconferencing, has made notable progress fueled by artificial intelligence (Chen et al. 2019; Zhou et al. 2020; Wang et al. 2021; Prajwal et al. 2020; Zhang et al. 2023; Yu et al. 2023; Tian et al. 2024; Xu et al. 2024; Drobyshev et al. 2024). However, various problems persist in existing animation methods. Specifically, some struggle with maintaining natural mouth shapes when animating exaggerated expressions (Zhou et al. 2020; Prajwal et al. 2020; Zhang et al. 2023), while others encounter severe warping distortions and identity alternations for unseen data (Wang et al. 2021; Ji et al. 2022). In addition to the difficulties in generating audio-synchronized lip motions, there are challenges in accurately estimating other spontaneous motions like head poses and eye motions, often resulting in poor performance (Zhou et al. 2020; Zhang et al. 2023) or reliance on another reference video (Zhou et al. 2021; Ji et al. 2022; Ma et al. 2023a,b), which is not available in most scenarios. Consequently, crafting a robust portrait animation framework that is effective for all types of input portraits and can independently generate satisfactory talking motions remains an unresolved issue.

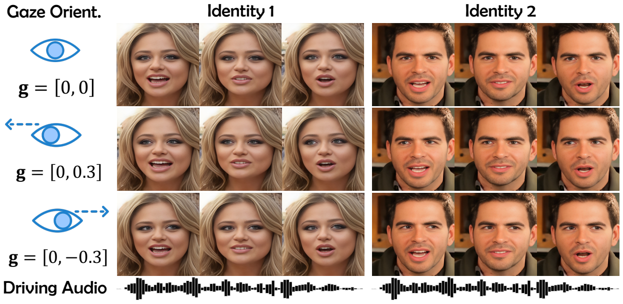


Figure 1: Illustration of gaze orientation experiments. The results of two identities driven by the same audio clip and different gaze directions are presented. The true pitch and yaw angles are multiplied by π.

Therefore, to devise a novel talking face system that can generalize well to any initial identities with various facial motions, several challenges remain to be addressed: 1) The input portraits may vary significantly in terms of appearance, expression, and other factors, requiring the system to learn a robust representation of facial features and movements that can be applied to new, unseen subjects. Animating techniques of current methods (Wang et al. 2021; Ji et al. 2022; Zhou et al. 2020) fail to fully disentangle identity and motion, resulting in poor generalization and distortions, especially when applied to out-of-distribution images with rich expressions. 2) Existing implicit or explicit motion representation methods (Deng et al. 2019; Zhou et al. 2021; Yu et al. 2023) for facial animation encounter limitations in gaze orientation, inevitably creating unnatural looking directions in the generated videos. 3) The intricate map-

[*]Corresponding author: qweizework@gmail.com.

pings between audio and low-frequency motions (e.g. head poses) require models to incorporate both prosody awareness and result diversity. Prior works that use probabilistic methods (Zhang et al. 2023) or sequence-to-sequence models (Wang et al. 2021) often emphasize one aspect—either diversity or prosody—while ignoring a balanced consideration of both features. 4) Learning precise lip-audio alignment requires enormous training sample pairs to achieve cross-modal adaption in the feature spaces, which is often inaccessible to regular researchers. Additionally, the interplay between mouth movements and other spontaneous facial actions (less correlated with audio), such as blinks and frowns, can introduce complexity to the overall expression generation process.

To overcome these difficulties, we propose **GoHD**, a **G**aze-**o**riented and **H**ighly **D**isentangled portrait animation method with audio-driven rhythmic head poses and realistic facial expressions. Specifically, GoHD is composed of three main modules: a generalized latent navigable face animator with gaze orientation, a prosody-aware denoising network for pose generation, and an expression estimator trained in a two-stage manner. Firstly, to accomplish fully disentangled motion transformation for arbitrary input identity, we integrate the animation module with latent navigation techniques (Wang et al. 2022b), skillfully decoupling a latent motion space from the underlying identity. More precisely, we split it into a source branch and a driving branch. In the source branch, a latent identity code is generated for each input reference image, representing the appearance feature without any head poses or expressions. Meanwhile, the driving branch processes target motions as inputs and predict a motion vector based on a learned motion codebook. Gaze directions are incorporated as conditions in this branch to provide overall motion control and rectify potential unnatural eye movements. The animated result is then obtained by decoding a combined representation of the predicted motion vector and the identity code.

Additionally, to realize audio-driven and controllable portrait animation, we design two independent generators for the driving motions in the face animator. An audio-conditioned diffusion model with a conformer-based denoising network is used to map audio cues to head poses, capturing prosody patterns with dilated convolutions and self-attention modules for natural, sequential results. The great probabilistic sampling characteristic (Ho, Jain, and Abbeel 2020; Alexanderson et al. 2023; Kong et al. 2021; Shen et al. 2023) of diffusion models further enhance the diversity of generated outputs. Regarding expressions, we focus on audio-related eye and lip motions, where lip movements require precise frame-wise synchronization, and eye motions like blinks and frowns depend more on temporal dynamics. To bridge this gap, we extract handcrafted eye motion features from pre-defined expression coefficients and introduce an audio-to-expression prediction approach trained by a two-stage strategy. The first stage focuses on distilling precise frame-wise lip motions from an expert pre-trained on sufficient audio-visual pairs (Prajwal et al. 2020), while the second stage uses an LSTM (Long Short-Term Memory) structured model to generate temporally dependent eye motions. With our well-designed two-stage training scheme, realistic and audio-synchronized expression generation is achieved with effective disentanglement of lip and eye motions.

In summary, this paper contributes in the following ways: 1) We propose a gaze-oriented and robust face animation module using latent navigation that effectively disentangles motion from identity. 2) We present a conformer-based conditional diffusion model for generating rhythmic and realistic poses. 3) A two-stage training strategy for expression prediction is devised to bridge the frequency gap between lip and eye motions. 4) Extensive experiments demonstrate that our method can generate advanced talking face results on arbitrary subjects with the proposed motion generation and animation modules.

2 Related Work

Audio-driven Talking Face Animation. The goal of this task is to generate a video where the input face image animates in synchronization with the provided audio. Early approaches (Chung et al. 2017; Vougioukas, Petridis, and Pantic 2019; Song et al. 2019) adopt end-to-end networks for direct frame-wise generation from input face image and audio. To enhance audio-visual control, Chen et al. (2019) uses explicit facial landmarks, while Zhou et al. (2019) employs disentangled latent representations. PC-AVS (Zhou et al. 2021) addresses spontaneous motions like head poses with a decoupled latent pose space. StyleTalk (Ma et al. 2023a) introduces a style-controllable decoder, and Yu et al. (Yu et al. 2023) decompose the latent space into lip and non-lip spaces. Some other works (Wang et al. 2021, 2022a) independently predict head motions but can lead to face distortion and identity alternation. MakeItTalk (Zhou et al. 2020) estimates speaker-specific motions with facial landmarks, limiting expression conveyance. MODA (Liu et al. 2023) enhances motion decoupling with denser landmarks. Later works (Ren et al. 2021; Zhang et al. 2021, 2023) explore 3DMMs, but appear desynchronized lip motions (Ren et al. 2021; Zhang et al. 2021) and unrealistic poses (Zhang et al. 2023). More recently, the world's famous AI labs released several outstanding works (He et al. 2023; Xu et al. 2024; Drobyshev et al. 2024; Tian et al. 2024) in this area, yet their requirements for huge training datasets are not practical to regular researchers. Our work introduces a novel framework capable of generating more realistic overall facial motions while addressing the practical challenge of limited training data availability.

Video-driven Talking Face Motion Imitation. In this category, the objective is to create a new video where the source face image adeptly mimics the expressions and head movements of the input driving video. Intermediaries are crucial for precise motion transformation. FOMM (Siarohin et al. 2019) uses learned key points and their affine transformations as structural references, while methods (Wang, Mallya, and Liu 2021; Siarohin et al. 2021; Hong et al. 2022; Zhao and Zhang 2022) enhance it with 3D (Wang, Mallya, and Liu 2021) or depth information (Hong et al. 2022), and modified motion estimation (Siarohin et al. 2021; Zhao and Zhang 2022). In contrast, LIA (Wang et al. 2022b) introduces a

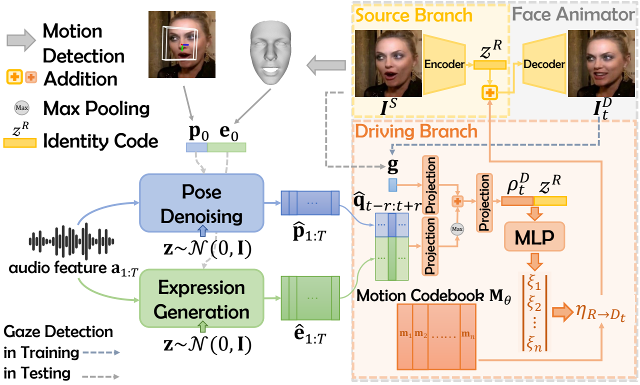


Figure 2: Illustration of our proposed **GoHD**, which is a highly disentangled and controllable taking face generation framework as described at the beginning of Section 3.

motion warping framework, navigating in latent space to avoid errors from explicit representations. Pang et al. (Pang et al. 2023) extend this approach with bidirectional cyclic training for disentangled pose and expression editing. StyleHEAT (Yin et al. 2022) uses a pre-trained StyleGAN (Karras et al. 2020) for high-resolution motion driving and editing, but may lead to identity discrepancies and artifacts. Our method adopts a latent navigable approach (Wang et al. 2022b) for simple and effective motion transformation in animating talking faces with predicted coefficients.

3 Method

The whole pipeline of our method is shown in Fig. 2. Given an audio clip with T frames of mel-spectrogram ($\mathbf{a}_{1:T}$) and an input source image $\boldsymbol{I}^S$, denoting its original pose and expression coefficients as $\mathbf{p}_0$ and $\mathbf{e}_0$ respectively, with a driving gaze direction $\mathbf{g}$ (detected from $\boldsymbol{I}^S$ or personally defined), the sequential talking face frames $\hat{\boldsymbol{I}}^D_{1:T}$ are generated as follows:

1) Diffused Head Poses. A rhythmic head pose sequence $\hat{\mathbf{p}}_{1:T}$ is synthesized through a probabilistic diffusion model conditioned on the input audio frames $\mathbf{a}_{1:T}$ and the original parameter $\mathbf{p}_0$.

2) Audio-to-expression Prediction. A sequence of expression coefficients, denoted as $\hat{\mathbf{e}}_{1:T}$, is obtained by a predictor trained in two stages, integrating an MLP-based (Multilayer Perceptron) distillation network and a generative LSTM model, from the given audio segment $\mathbf{a}_{1:T}$ and the original expression parameter $\mathbf{e}_0$.

3) Gaze-oriented Face Animation. Given the predicted motion descriptors $\hat{\mathbf{p}}_{1:T}$ and $\hat{\mathbf{e}}_{1:T}$, and the predetermined gaze orientation $\mathbf{g}$, the input source image $\boldsymbol{I}^S$ can be animated to $\hat{\boldsymbol{I}}^D_{1:T}$ frame by frame through an animation module involving latent space navigation to achieve robust facial motion transformations.

3.1 Diffusion-based Head Pose Generator

Diffusion Model. To ensure that the generated head motions exhibit diversity while maintaining a sense of rhythmicity,

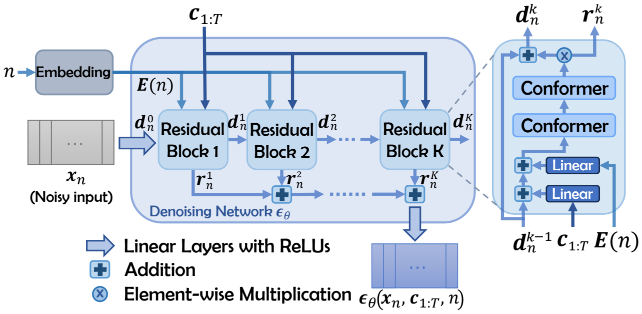


Figure 3: Demonstration of the residual denoising network architecture in the diffusion model for head pose estimation.

we design a conditional diffusion model to synthesis head pose coefficients $\hat{\mathbf{p}}_{1:T}$ corresponding to the input audio feature $\mathbf{a}_{1:T}$. The diffusion process at step $n \in \{1, 2, ..., \mathbf{N}\}$ is defined as follows:

$$q(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}) = \mathcal{N}(\boldsymbol{x}_n; \sqrt{\alpha_n}\boldsymbol{x}_{n-1}, \beta_n\mathbf{I}), \tag{1}$$

where $\alpha_n = 1 - \beta_n (0 < \beta_n < 1)$ (2020), so that $\{\beta_n\}_{n=1}^{\mathbf{N}}$ completely defines the diffusion process by sampling $\boldsymbol{x}_n = \sqrt{1-\beta_n}\boldsymbol{x}_{n-1} + \sqrt{\beta_n}\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. In our context, we designate the original variable $\boldsymbol{x}_0$ as the residual sequence of the ground-truth pose coefficients $\Delta\mathbf{p}_{1:T} = \mathbf{p}_{1:T} - \mathbf{p}_0 \in \mathbb{R}^{T\times 6}$ to generate more natural and continuous motion over the first pose of the sequence.

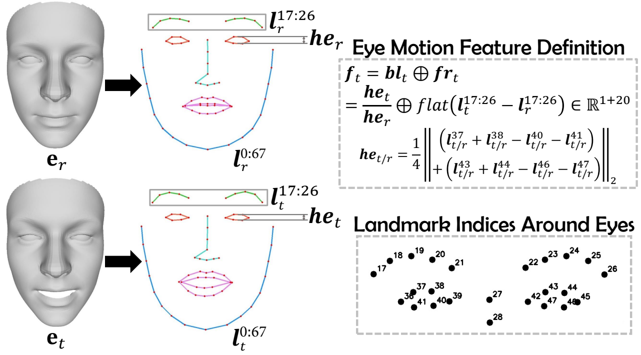
As formulated by DDPM (2020), the network only needs to predict the added noise $\boldsymbol{\epsilon}$, thus the loss function can be constructed as:

$$\mathcal{L}(\theta|\mathcal{D}) = \mathbb{E}_{\boldsymbol{x}_0,n,\boldsymbol{\epsilon}}\big[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_n}\boldsymbol{x}_0 + \sqrt{\bar{\beta}_n}\boldsymbol{\epsilon}, \boldsymbol{c}, n)\|^2\big], \tag{2}$$

where $\boldsymbol{x}_0$ is uniformly sampled from the training data $\mathcal{D}$, $\bar{\alpha}_n$ and $\bar{\beta}_n$ are constants determined by $\{\beta_n\}_{n=1}^{\mathbf{N}}$, and $\boldsymbol{\epsilon}_\theta$ is the conditional noise prediction network with learnable parameters. In our case, the condition variable $\boldsymbol{c}$ can either be the input audio feature $\mathbf{a}_{1:T}$ or its combination with the initial pose coefficient $\mathbf{p}_0$.

Network Architecture. The architecture of our denoising network is shown in Fig. 3. Inspired by (Kong et al. 2021), we implement the network with a series of conditional residual blocks for generating audio-aware residual pose sequences. Within each block, we stack two conformers where attention modules are incorporated into dilated convolutions to effectively assimilate information over extended time scales.

Head Pose Synthesis. To enhance the realism of the synthesized results, we incorporate classifier-free guidance (Ho and Salimans 2021) to partially condition the reverse diffusion process on the initial pose $\mathbf{p}_0$. Given an input reference pose $\mathbf{p}_0$, we define the source-referred conditioning $\boldsymbol{c}_{1:T}$ with $\boldsymbol{c}_t = \mathbf{a}_t \oplus \mathbf{p}_0$, where $\mathbf{a}_t$ is the audio feature at the t-th frame. After separately training a $\mathbf{p}_0$-conditional model $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_n, \boldsymbol{c}_{1:T}, n)$ and a $\mathbf{p}_0$-unconditional model $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_n, \mathbf{a}_{1:T}, n)$, the classifier-free guidance can be



$\mathbf{e}_r$
$\boldsymbol{l}_r^{17:26}$
$\boldsymbol{he}_r$
$\boldsymbol{l}_r^{0:67}$
$\mathbf{e}_t$
$\boldsymbol{l}_t^{17:26}$
$\boldsymbol{he}_t$
$\boldsymbol{l}_t^{0:67}$
Eye Motion Feature Definition
$\boldsymbol{f}_t = \boldsymbol{bl}_t \oplus \boldsymbol{fr}_t$
$= \frac{\boldsymbol{he}_t}{\boldsymbol{he}_r} \oplus flat(\boldsymbol{l}_t^{17:26} - \boldsymbol{l}_r^{17:26}) \in \mathbb{R}^{1+20}$
$\boldsymbol{he}_{t/r} = \frac{1}{4}\left\| \begin{matrix} (\boldsymbol{l}_{t/r}^{37} + \boldsymbol{l}_{t/r}^{38} - \boldsymbol{l}_{t/r}^{40} - \boldsymbol{l}_{t/r}^{41}) \\ + (\boldsymbol{l}_{t/r}^{43} + \boldsymbol{l}_{t/r}^{44} - \boldsymbol{l}_{t/r}^{46} - \boldsymbol{l}_{t/r}^{47}) \end{matrix} \right\|_2$
Landmark Indices Around Eyes
17
18
19
20
21
22
23
24
25
26
36
37
38
39
40
41
27
28
42
43
44
45
46
47

Figure 4: Definition of the eye motion feature, where $\boldsymbol{bl}_t$ represents the eye-blinking ratio of the t-th frame, with $\boldsymbol{he}_{t/r} \in \mathbb{R}$ denoting the average heights of eyes. $\boldsymbol{fr}_t \in \mathbb{R}^{20}$ symbolizes the corresponding brow displacements, and $flat$ means the operation of flattening. The landmark indices and calculation for $\boldsymbol{he}_{t/r}$ are illustrated on the right side.

achieved by combining the prediction of both models:

$$\begin{aligned}\boldsymbol{\epsilon}_\theta^\gamma(\boldsymbol{x}_n, \mathbf{c}_{1:T}, n) =& \gamma\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_n, \mathbf{c}_{1:T}, n) \\ &+ (1-\gamma)\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_n, \mathbf{a}_{1:T}, n),\end{aligned} \tag{3}$$

where the coefficient $\gamma (0 < \gamma < 1)$ can be adjusted to control the influence of $\mathbf{p}_0$, and the synthesized head pose sequence can be inferred by:

$$\hat{\mathbf{p}}_{1:T} = \mathbf{p}_0 + \hat{\Delta}\mathbf{p}_{1:T} = \mathbf{p}_0 + \hat{\boldsymbol{x}}_0. \tag{4}$$

where $\hat{\boldsymbol{x}}_0$ is the reverse sampling result given the predicted noise.

3.2 Expression Predictor Trained in Two Stages

This module includes frame-wise distillation for audio-synchronized mouth shapes and temporal prediction for spontaneous eye motions. To facilitate the disentanglement of various facial actions from the overall expression coefficients, we pre-define handcrafted eye motion features as control signals for the first stage and as the generation goal of the second stage.

Handcrafted Eye Motion Features. Given the expression coefficients $\mathbf{e}_t$ of the t-th frame, the facial mesh can be reconstructed by setting all other coefficients (poses and identity) to zero. We then extract the 68 facial landmarks $\boldsymbol{l}_t^{0:67}$ from the above mesh. Similarly, a reference set of landmarks $\boldsymbol{l}_r^{0:67}$ is obtained from a neutral facial mesh with "mean expression" $\mathbf{e}_r$, where all coefficients of the expression basis are set to zero. Considering eye blinks and brow frowns, we define the eye motion feature $\boldsymbol{f}_t \in \mathbb{R}^{21}$ in Fig. 4.

Stage 1: Audio-to-lip Distillation. The audio-to-lip mapping poses a one-to-one problem due to the strong connection between mouth shape and pronunciation. To ensure that the network specifically learns the correlation between audio and lip motions in the first stage, we incorporate the ground-truth handcrafted eye motion features $\boldsymbol{f}_{1:T}$ as additional input signals along with the audio $\mathbf{a}_{1:T}$ and the initial expression coefficients $\mathbf{e}_0$ for regressing the overall expressions. The mapping of each frame can be written as:

$$\tilde{\mathbf{e}}_t = \mathbf{MLP}_\theta(\boldsymbol{\Phi}_a(\mathbf{a}_t) \oplus \mathbf{e}_0 \oplus \boldsymbol{f}_t), \tag{5}$$

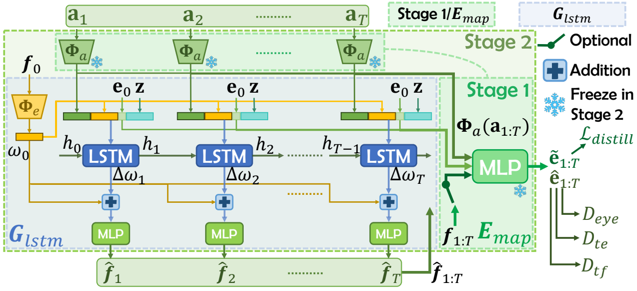


Figure 5: The expression predictor trained in two stages.

where $\boldsymbol{\Phi}_a$ is an audio encoder that embeds the input audio feature to a latent space and $\mathbf{MLP}_\theta$ denotes a multilayer perception. Notably, we distill the resynchronized results from a pre-trained lip expert (Prajwal et al. 2020) ($\mathcal{L}_{distill}$) to inherit its lip-audio alignment capability learned on sufficient sample pairs, thereby compensating for our limited dataset and reducing the risk of under-fitting.

Stage 2: Eye Motion Generation. After learning synchronized audio-to-lip mapping, the second stage focuses on addressing the more complex mapping between audio and eye motions and is trained with the learned weights in stage 1 frozen. To tackle this generation problem containing more temporal dynamics, we employ the LSTM architecture. This choice over transformer models is deliberate, as LSTMs are known for their robustness in handling longer sequences during inference, ensuring effective modeling of information dependencies across various time scales. As is depicted in Fig. 5, taking the sequence of audio features $\mathbf{a}_{1:T}$, the initial expression coefficients $\mathbf{e}_0$, and the corresponding eye motion feature $\boldsymbol{f}_0$ as input, we first encode $\mathbf{a}_{1:T}$ through the audio encoder $\boldsymbol{\Phi}_a$ pre-trained in the first stage. Along with $\mathbf{e}_0$ and the encoded eye motion features $\omega_0 = \boldsymbol{\Phi}_e(\boldsymbol{f}_0)$, the sequential procedure can be described as follows:

$$(h_t, \Delta\omega_t) = \mathbf{LSTM}_\theta(h_{t-1}, \boldsymbol{\Phi}_a(\mathbf{a}_t) \oplus \omega_0 \oplus \mathbf{e}_0 \oplus \mathbf{z}), \tag{6}$$

$$\hat{\boldsymbol{f}}_t = \mathbf{MLP}_\theta(\omega_t) = \mathbf{MLP}_\theta(\omega_0 + \Delta\omega_t), \tag{7}$$

where h_t is the hidden state at time step t, which corresponds to the t-th frame, while h_0 is a zero vector with the same shape as ω_0. To encourage the network to learn multiple probabilities of generating spontaneous motions, we also concatenated a latent vector $\mathbf{z}$ in the hidden layer at each step, which is sampled from the standard multivariate Gaussian distribution. Note that we have the network predict the residuals of the embedded eye motion features for faster convergence and better generalization ability. Combining with the well-trained mapping network from the first stage, the overall estimation of audio-driven expressions is completed in a single, cohesive process:

$$\begin{aligned}\hat{\mathbf{e}}_{1:T} &= \boldsymbol{E}_{map}(\mathbf{a}_{1:T}, \mathbf{e}_0, \hat{\boldsymbol{f}}_{1:T}) \\ &= \boldsymbol{E}_{map}(\mathbf{a}_{1:T}, \mathbf{e}_0, \boldsymbol{G}_{lstm}(\boldsymbol{\Phi}_a(\mathbf{a}_{1:T}), \boldsymbol{f}_0, \mathbf{e}_0, \mathbf{z})),\end{aligned} \tag{8}$$

where $\boldsymbol{E}_{map}$ and $\boldsymbol{G}_{lstm}$ are the frame-wise mapping network and the LSTM-based eye motions generator, respectively. We introduce three discriminators $\boldsymbol{D}_{eye}$, $\boldsymbol{D}_{te}$ and $\boldsymbol{D}_{tf}$ to help distinguish the temporal naturalness and realness of the results. Extended descriptions can be found in the *supplementary material.*

3.3 Latent Navigable Face Animator

Given the pose and expression coefficients $\hat{\mathbf{p}}_{1:T}$ and $\hat{\mathbf{e}}_{1:T}$ predicted from the audio, along with the reference image $\boldsymbol{I}^S$ and target gaze direction $\mathbf{g}$, we draw inspiration from (Wang et al. 2022b) and introduce a well-designed face animator to generate the final talking portrait frames $\hat{\boldsymbol{I}}^D_{1:T}$. Unlike previous methods that rely on the transformations of spatial key points (Wang, Mallya, and Liu 2021; Siarohin et al. 2019), our animator directly manipulates the latent space to alleviate information loss caused by using explicit structural representations and achieve better disentanglement of identity and motion. Additionally, different from (Wang et al. 2022b) that requires a real video as the overall motion-driving signal, our animator makes the animation derivable through separate intermediate motion descriptors. This design choice enables explicit editing of various facial attributes and supports multi-modal driving (Fig. 9). In training time, it animates the source image in a frame-by-frame manner by learning the motion transformation from $\boldsymbol{I}^S$ to the target *t*-th frame $\boldsymbol{I}^D_t$ via the detected coefficients $\mathbf{q}_{t-r:t+r} = \mathbf{p}_{t-r:t+r} \oplus \mathbf{e}_{t-r:t+r}$, where $\boldsymbol{I}^S$ and $\boldsymbol{I}^D_t$ are two randomly selected frames of a video, and r is the radius of the adjacent window for smoothing, which is achieved by a max pooling layer after several layers of projection. As shown in Fig. 2, we first encode the source image into a latent space to acquire an identity code z^R. This latent vector is then concatenated with the projected and gaze-conditioned driving feature ρ^D_t to estimate the motion transfer $\eta_{R \to D_t}$ on a learnable motion codebook, which consists of a series of learnable orthogonal motion directions $\mathbf{M}_\theta = \{\mathbf{m}_1, ...\mathbf{m}_n\}$ to represent any latent navigation. By jointly learning the magnitude ξ_i of each direction $\mathbf{m}_i$, the latent navigation can be linearly calculated as follows:

$$\eta_{R \to D_t} = \sum_{i \in [1,n]} \xi_i \mathbf{m}_i, \tag{9}$$

where $\xi_{1:n} = \mathbf{MLP}_\theta(\rho^D_t \oplus z^R)$. Afterward, the target latent representation can be obtained by simple addition: $z^D_t = z^R + \eta_{R \to D_t}$, from which the target frame $\boldsymbol{I}^D_t$ will be generated through a decoder. Notice that, during training, the driving gaze directions are directly inherited from the driving frames, then the module's gaze orientation ability can be optimized through a simple gaze loss. During inference, the driving gaze directions can be set to any reasonable pitch and yaw angles to achieve effective gaze manipulation or rectify potentially unnatural looking directions. For simplicity, we set them to the original gaze directions derived from $\boldsymbol{I}^S$s in most of our experiments.

4 Experiments

4.1 Experimental Settings

Datasets. We leverage a subset of the VoxCeleb dataset (2017) as the training set for our face animator, and a portion of the HDTF dataset (2021) for the generation of motion descriptors. Most of the testing is also conducted on hundreds of unseen videos from these two datasets.

Method	HDTF			VoxCeleb		
	LSE-C ↑	LSE-D ↓	FID ↓	LSE-C ↑	LSE-D ↓	FID ↓
MakeItTalk (2020)	5.58	9.85	34.28	4.40	10.42	61.98
Wav2Lip (2020)	**10.04**	**5.93**	34.40	**9.32**	**6.15**	67.75
Audio2Head (2021)	7.97	$\underline{7.30}$	34.31	5.79	8.61	79.05
EAMM (2022)	5.45	9.57	57.34	4.74	9.61	85.39
SadTalker (2023)	7.60	7.70	36.91	6.99	7.75	60.75
Ours	$\underline{8.13}$	7.78	**30.40**	$\underline{7.20}$	$\underline{7.70}$	**58.11**
Ground Truth	8.97	6.67	-	7.51	7.42	-

Table 1: Quantitative comparisons for lip synchronization and video quality. Best results are in **bold**, and scores closest to the ground truth are underlined for reference.

Dataset	Method / Metrics	Audio2Head	SadTalker	Ours	Ground Truth
HDTF	$\mathrm{Var}_p^{\times 10^3}$	2.399	$\underline{2.473}$	2.411	4.315
	SSIM_p ↑	0.972	0.985	**0.996**	-
	$\mathrm{Var}_e^{\times 10^2}$	3.066	2.206	$\underline{5.374}$	9.778
	SSIM_e ↑	0.819	0.904	**0.915**	-
VoxCeleb	$\mathrm{Var}_p^{\times 10^3}$	1.992	1.896	$\underline{2.314}$	8.746
	SSIM_p ↑	0.984	**0.987**	**0.987**	-
	$\mathrm{Var}_e^{\times 10}$	0.618	0.182	$\underline{0.865}$	1.585
	SSIM_e ↑	0.754	0.854	**0.872**	-

Table 2: Quantitative comparisons for spontaneous motions. Best SSIM scores are in **bold**, and variances closest to the ground truth are underlined.

Comparison Methods. We conduct a comprehensive evaluation of our method by comparing it with various advanced audio-only driven methods, including Wav2Lip (Prajwal et al. 2020), MakeItTalk (Zhou et al. 2020), Audio2Head (Wang et al. 2021), EAMM (Ji et al. 2022), and SadTalker (Zhang et al. 2023). Our evaluation covers lip synchronization and video quality for all the mentioned approaches. Additionally, we assess the data structural match and naturalness of other spontaneous motions when compared to SadTalker and Audio2Head.

Evaluation Metrics. We use Frechet Inception Distance (FID) (Heusel et al. 2017) to evaluate image quality. For lip synchronization, we adopt methods from previous works (Zhang et al. 2023; Yu et al. 2023) and utilize the pre-trained SyncNet (Prajwal et al. 2020) for confidence (LSE-C) and distance (LSE-D) evaluations of lip motions. Using the 2D landmarks derived from the detected expression coefficients, we compute the structural similarity (SSIM_e) and average variance (Var_e) on the eyes and brows landmarks sequences to assess the naturalness of eye motions. On the other hand, to evaluate poses, we employ a pre-trained pose detection model (Algabri, Shin, and Lee 2024) to obtain the pose sequences of the generated videos. We then calculate the structural similarities between these sequences (SSIM_p) and the average variance of their corresponding feature vectors (Var_p) to indicate their statistical match with real data.

4.2 Comparison with State-of-the-art Methods

Quantitative Comparison. Quantitative evaluations for lip synchronization and video quality are reported in Table 1.

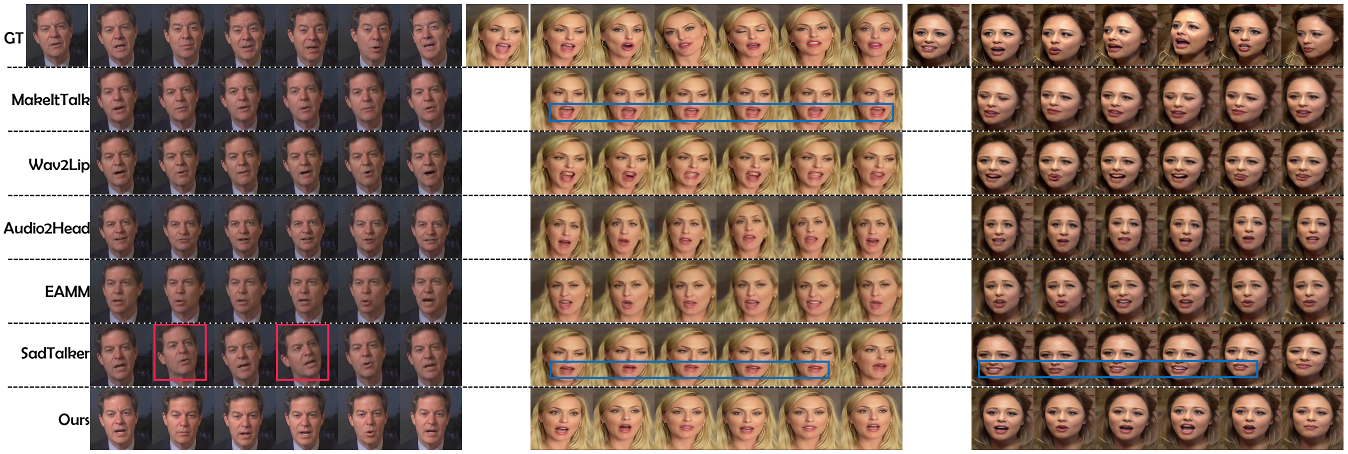


Figure 6: Qualitative comparison on the two datasets. Apart from accurate lip synchronization, our method presents the best generalization capability on animating extravagant input expressions and stability in pose generation.

According to the FID, our approach demonstrates an overall improvement in the realism of generated videos. In terms of lip-sync performance, Wav2Lip unquestionably achieves the best results, surpassing even the ground truth, because it directly trains with the SyncNet model used for evaluation. Consequently, we interpret scores closer to the ground truth as indicating a relatively better ability to produce realistic mouth movements. In this context, our method exhibits better performance than SadTalker. Meanwhile, Audio2Head presents smaller lip motion distances on the HDTF dataset, likely due to the overlap between our testing set and its training set. Furthermore, Table 2 illustrates assessments for spontaneous motions. Two representative methods (Audio2Head and SadTalker) are included in this comparison. Audio2Head exhibits high diversity in generated poses and expressions but suffers from significant misalignment with real data, especially in expressions. In contrast, SadTalker demonstrates good structural similarity with the ground truth, albeit with lower diversity, especially in eye motions, as it only considers controllable blinks in expression generation. Our GoHD achieves a balance between data diversity and realism, presenting a comprehensive advancement in poses and eye motion generation.

Qualitative Comparison. Fig. 6 shows visual comparisons of three examples from the HDTF and the VoxCeleb dataset. Audio2Head (Wang et al. 2021) and EAMM (Ji et al. 2022) both rely on the animation framework of FOMM (Siarohin et al. 2019), exhibiting severe face distortions and struggling to preserve identity. MakeItTalk (Zhou et al. 2020) performs poorly in lip synchronization, while Wav2Lip (Prajwal et al. 2020) suffers from artifacts in the lip region, especially when handling substantial variations in mouth shape. SadTalker demonstrates relatively high visual quality but occasionally produces unnatural and upward-tilted head poses. As shown on the middle and the right, it encounters incomplete motion disentanglement and has difficulty animating faces with exaggerated lip morphology. In general, aside from accurate lip synchronization, our method demonstrates superior generalization capability in animating extravagant input expressions and stability in pose generation.

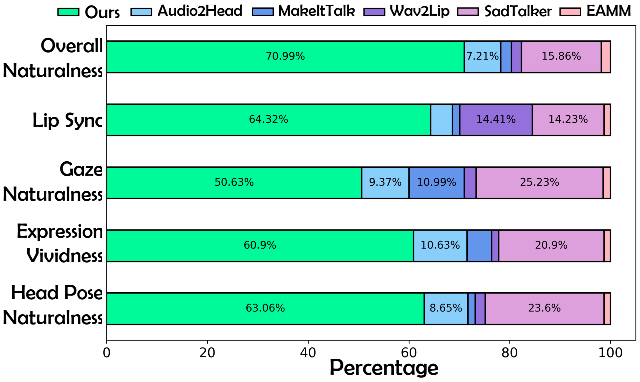


Figure 7: The result of user study.

User Study. We conduct a user study to evaluate the overall performance of our method against various competitors. We randomly select 30 test examples and invite 37 volunteers to assess each example in terms of head pose naturalness, expression vividness (with a focus on eye motions like blinks and frowns), gaze naturalness, lip synchronization, and overall naturalness. With a total of $30 \times 37 = 1,110$ responses for each attribute, the support percentages for each method are depicted in Fig. 7. Notably, our method outperforms all others on comprehensive aspects, receiving 70.99% of the responses for overall naturalness.

4.3 Validation Experiments

Motion Interpolation. We provide visualizations for motion interpolation of our face animator to showcase its robustness in motion editing. The reference images $\boldsymbol{I}^R$s, decoded from the reference latent representations z^Rs, consistently exhibit a frontal pose and mean expression, demonstrating the effective disentanglement of motion from identity. In Fig. 8, as the coefficient λ of the latent navigation vector $\eta_{R\rightarrow D_t}$ linearly increases, the final image derived from $z^R + \lambda\eta_{R\rightarrow D_t}$ (denoted by $\lambda\eta_{R\rightarrow D_t}, \lambda \in 0.25, 0.5, 0.75$ in the figure) gradually transfers in all mo-

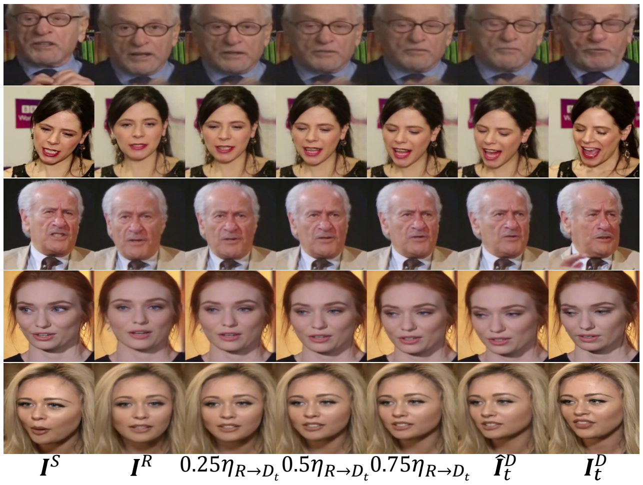


Figure 8: Visualization of motion interpolation.

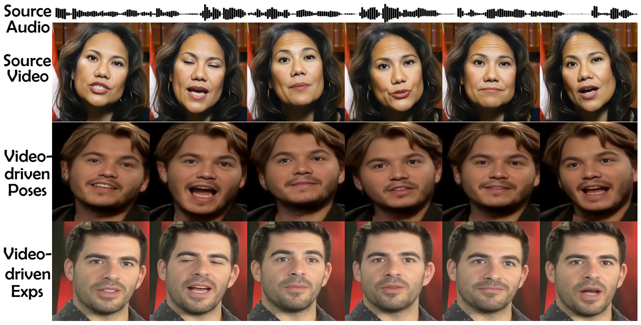


Figure 9: Demonstration of multi-modal driving results.

tions, until $\lambda = 1$ to reach the target ones, indicating the effectiveness and versatility of our method in controllable motion transformation and identity disentanglement.

Gaze Orientation. To validate the gaze control capability of our face animator, we set the yaw angle in the driving gaze direction to three different values $(0, 0.3\pi, -0.3\pi)$, corresponding to looking forward, left, and right, respectively. Using the same audio clip, the results for two identities are shown in Fig. 1, demonstrating effective manipulation on gaze orientation across various input portraits.

Multi-modal Driving. In addition to solely relying on audio-driven generation, our approach allows for the extraction of intermediate motion descriptors from a source video, enabling multimodal-driven animation. As illustrated in Fig. 9, the "Video-driven Exps" scenario involves deriving $\mathbf{e}_{1:T}$ from the source video and the $\mathbf{p}_{1:T}$ from the source audio, and vice versa. The video-driven signals in the results align with those of the source video, while allowing variations in the audio-driven component. These results demonstrate the effectiveness of our approach in achieving disentangled control over facial animations, with promising implications for multi-modal applications.

4.4 Ablation Study

Two-stage Strategy. To verify the effectiveness of our carefully designed two-stage expression predictor, we conduct ablation studies on 100 videos of the HDTF dataset with the following variants: 1) w/o **Stage 2**&$\boldsymbol{f}_{1:T}$: Produce the expression coefficients in a regressive manner by only employing the mapping network in Stage 1 without inputting eye motions features. 2) w/o **Stage 1**: Generate expressions directly through Stage 2 without pre-training the Stage 1 network. 3) w/o Distillation: Use the ground-truth lip motions as the training target in Stage 1 instead of distilling from the lip expert. 4) **Full**-transformer: Our full training strategy with $\boldsymbol{G}_{lstm}$ replaced by a transformer model. 5) **Full**-LSTM: Our full training strategy.

Strategy	MLD ↓	SSIM$_e$ ↑
w/o **Stage 2**&$\boldsymbol{f}_{1:T}$	**1.785**	0.813
w/o **Stage 1**	1.931	0.836
w/o Distillation	2.012	0.852
Full-transformer	1.806	0.894
Full-LSTM	1.792	**0.915**

Table 3: Ablation results for the two-stage strategy in expression prediction. The best results are highlighted in **bold**.

We compute the average mouth landmark distances (MLD) and eye motion structural similarities (SSIM$_e$) for the generated expression coefficients to evaluate each design choice in lip synchronization and eye motion generation. The numerical results are reported in Table 3. The **Full**-strategies demonstrate enhanced alignment of mouth shapes compared to the variant w/o **Stage 1**. This underscores the significant role played by the pre-trained first stage in learning lip motions synchronized with audio, where the distillation approach is also indispensable. Despite achieving the best performance in lip synchronization, employing only a mapping network to predict expressions (w/o **Stage 2**&$\boldsymbol{f}_{1:T}$) faces challenges in producing realistic eye motions and leads to poor SSIM$_e$ score. In contrast, our **Full**- models simultaneously achieve higher lip-sync quality and naturalness in the results. Notably, the LSTM-based architecture surpasses the transformer-based one due to its ability to effectively model dependencies between neighboring frames, contributing to the overall enhanced performance in lip-sync generation by enabling more accurate sequential prediction.

5 Conclusion

In this work, we introduce **GoHD**, a novel and robust framework for generating realistic audio-driven talking faces. Beyond pose and expression coefficients, we incorporate gaze direction as an additional driving condition for gaze-oriented animation. We employ a conformer-structured conditional diffusion model to synthesize rhythmic head poses. For audio-driven expression generation, we devise a predictor trained in a two-stage manner that separates frame-wise and frequent lip motions from other temporally dependent but less audio-related movements. Moreover, a latent navigable animation module is proposed for gaze-oriented and robust motion transformation. Experimental results illustrate the superiority of our GoHD to produce high-quality talking videos for any subject.

Acknowledgments

This work was partially funded by the National Natural Science Foundation of China (62102418, 61932003, 62372026, 62172415), the Beijing Science and Technology Plan Project (Z231100005923033), and the Excellent Youth Program of State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS2024312).

References

Alexanderson, S.; Nagy, R.; Beskow, J.; and Henter, G. E. 2023. Listen, Denoise, Action! Audio-Driven Motion Synthesis with Diffusion Models. *ACM Trans. on Graphics (TOG)*, 42(4): 44:1–44:20.

Algabri, R.; Shin, H.; and Lee, S. 2024. Real-time 6DoF full-range markerless head pose estimation. *Expert Systems with Applications*, 239: 122293.

Chen, L.; Maddox, R. K.; Duan, Z.; and Xu, C. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 7832–7841.

Chung, J.; Jamaludin, A.; Zisserman, A.; et al. 2017. You said that? In *British Machine Vision Conference (BMVC)*. British Machine Vision Association and Society for Pattern Recognition.

Deng, Y.; Yang, J.; Xu, S.; Chen, D.; Jia, Y.; and Tong, X. 2019. Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Drobyshev, N.; Casademunt, A. B.; Vougioukas, K.; Landgraf, Z.; Petridis, S.; and Pantic, M. 2024. EMOPortraits: Emotion-enhanced Multimodal One-shot Head Avatars. arXiv:2404.19110.

He, T.; Guo, J.; Yu, R.; Wang, Y.; Zhu, J.; An, K.; Li, L.; Tan, X.; Wang, C.; Wu, H.; Zhao, S.; and Bian, J. 2023. GAIA: Zero-shot Talking Avatar Generation. In *ICLR 2024*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 6629–6640. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. 33: 6840–6851.

Ho, J.; and Salimans, T. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Hong, F.-T.; Zhang, L.; Shen, L.; and Xu, D. 2022. Depth-Aware Generative Adversarial Network for Talking Head Video Generation.

Ji, X.; Zhou, H.; Wang, K.; Wu, Q.; Wu, W.; Xu, F.; and Cao, X. 2022. EAMM: One-Shot Emotional Talking Face via Audio-Based Emotion-Aware Motion Model. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 8110–8119.

Kong, Z.; Ping, W.; Huang, J.; Zhao, K.; and Catanzaro, B. 2021. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *International Conference on Learning Representations (ICLR)*.

Liu, Y.; Lin, L.; Fei, Y.; Changyin, Z.; and Yu, L. 2023. MODA: Mapping-Once Audio-driven Portrait Animation with Dual Attentions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Ma, Y.; Wang, S.; Hu, Z.; Fan, C.; Lv, T.; Ding, Y.; Deng, Z.; and Yu, X. 2023a. StyleTalk: One-Shot Talking Head Generation with Controllable Speaking Styles. AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Ma, Y.; Zhang, S.; Wang, J.; Wang, X.; Zhang, Y.; and Deng, Z. 2023b. DreamTalk: When Expressive Talking Head Generation Meets Diffusion Probabilistic Models. *arXiv preprint arXiv:2312.09767*.

Nagrani, A.; Chung, J.; and Zisserman, A. 2017. VoxCeleb: a large-scale speaker identification dataset. *Interspeech*.

Pang, Y.; Zhang, Y.; Quan, W.; Fan, Y.; Cun, X.; Shan, Y.; and Yan, D.-M. 2023. DPE: Disentanglement of Pose and Expression for General Video Portrait Editing. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 427–436.

Prajwal, K. R.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. MM '20, 484–492. Association for Computing Machinery.

Ren, Y.; Li, G.; Chen, Y.; Li, T. H.; and Liu, S. 2021. PIRenderer: Controllable Portrait Image Generation via Semantic Neural Rendering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 13759–13768.

Shen, S.; Zhao, W.; Meng, Z.; Li, W.; Zhu, Z.; Zhou, J.; and Lu, J. 2023. DiffTalk: Crafting Diffusion Models for Generalized Audio-Driven Portraits Animation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019. First Order Motion Model for Image Animation.

Siarohin, A.; Woodford, O.; Ren, J.; Chai, M.; and Tulyakov, S. 2021. Motion Representations for Articulated Animation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Song, Y.; Zhu, J.; Li, D.; Wang, A.; and Qi, H. 2019. Talking Face Generation by Conditional Recurrent Adversarial Network. 919–925. International Joint Conferences on Artificial Intelligence Organization.

Tian, L.; Wang, Q.; Zhang, B.; and Bo, L. 2024. EMO: Emote Portrait Alive – Generating Expressive Portrait Videos with Audio2Video Diffusion Model under Weak Conditions. arXiv:2402.17485.

Vougioukas, K.; Petridis, S.; and Pantic, M. 2019. End-to-End Speech-Driven Realistic Facial Animation with Temporal GANs. In *Proc. of the IEEE International Conference on Computer Vision Workshops*.

Wang, S.; Li, L.; Ding, Y.; Fan, C.; and Yu, X. 2021. Audio2Head: Audio-driven One-shot Talking-head Generation with Natural Head Motion.

Wang, S.; Li, L.; Ding, Y.; and Yu, X. 2022a. One-shot Talking Face Generation from Single-speaker Audio-Visual Correlation Learning.

Wang, T.-C.; Mallya, A.; and Liu, M.-Y. 2021. One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Y.; Yang, D.; Bremond, F.; and Dantcheva, A. 2022b. Latent Image Animator: Learning to Animate Images via Latent Space Navigation. In *International Conference on Learning Representations (ICLR)*.

Xu, S.; Chen, G.; Guo, Y.-X.; Yang, J.; Li, C.; Zang, Z.; Zhang, Y.; Tong, X.; and Guo, B. 2024. VASA-1: Lifelike Audio-Driven Talking Faces Generated in Real Time. arXiv:2404.10667.

Yin, F.; Zhang, Y.; Cun, X.; Cao, M.; Fan, Y.; Wang, X.; Bai, Q.; Wu, B.; Wang, J.; and Yang, Y. 2022. Styleheat: One-shot high-resolution editable talking face generation via pretrained stylegan. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 85–101. Springer.

Yu, Z.; Yin, Z.; Zhou, D.; Wang, D.; Wong, F.; and Wang, B. 2023. Talking Head Generation with Probabilistic Audio-to-Visual Diffusion Priors. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Zhang, W.; Cun, X.; Wang, X.; Zhang, Y.; Shen, X.; Guo, Y.; Shan, Y.; and Wang, F. 2023. SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 8652–8661.

Zhang, Z.; Li, L.; Ding, Y.; and Fan, C. 2021. Flow-Guided One-Shot Talking Face Generation With a High-Resolution Audio-Visual Dataset. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 3661–3670.

Zhao, J.; and Zhang, H. 2022. Thin-plate spline motion model for image animation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 3657–3666.

Zhou, H.; Liu, Y.; Liu, Z.; Luo, P.; and Wang, X. 2019. Talking Face Generation by Adversarially Disentangled Audio-Visual Representation.

Zhou, H.; Sun, Y.; Wu, W.; Loy, C. C.; Wang, X.; and Liu, Z. 2021. Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

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