Emotion-Conditioned Motion Sub-spaces with Flow Matching for Real-Time Audio-Driven Talking Heads

Haoyu Wang[1], Xiaozhe Xin[3], Xiaoyu Qin[1]*, Meiguang Jin[3], Junfeng Ma[3], Dan Xu[4], Jia Jia[1,2,5]*

[1]Department of Computer Science and Technology, Tsinghua University,
[2]Key Laboratory of Pervasive Computing, Ministry of Education
[3]Alibaba Group,
[4]Hong Kong University of Science and Technology,
[5]BNRist, Tsinghua University,
why22@mails.tsinghua.edu.cn, xyqin@tsinghua.edu.cn, jjia@tsinghua.edu.cn

Abstract

Recent advances in audio-driven talking-head synthesis have brought lip-sync precision close to human perception, yet emotional fidelity and real-time inference remain open challenges. Existing pipelines typically disentangle lip articulation, facial expression, and head pose in latent space; this rigid factorization ignores the intrinsic coupling between articulation and affect — e.g., downward lip corners when sad—thus limiting expressiveness. We cast speech-conditioned facial motion as a sample from an emotion-conditioned distribution in a motion latent space. Concretely, we (i) learn a motion dictionary of orthogonal bases with an autoencoder via self-supervision, (ii) construct emotion-conditioned sub-spaces within the latent space, and (iii) design a layer-progressive cross-attention fusion module that modulates a flow-matching sampler with both audio and emotion signals. Only ten reverse ODE steps are required to generate a motion-latent trajectory, enabling real-time end-to-end latency. Extensive experiments on MEAD and RAVDESS show that our method outperforms recent GAN- and diffusion-based baselines in emotion accuracy while running at around 75 FPS on a single desktop GPU. The proposed framework delivers the first emotionally expressive Audio2Face system that simultaneously achieves lip-sync accuracy, affective realism, and real-time performance.

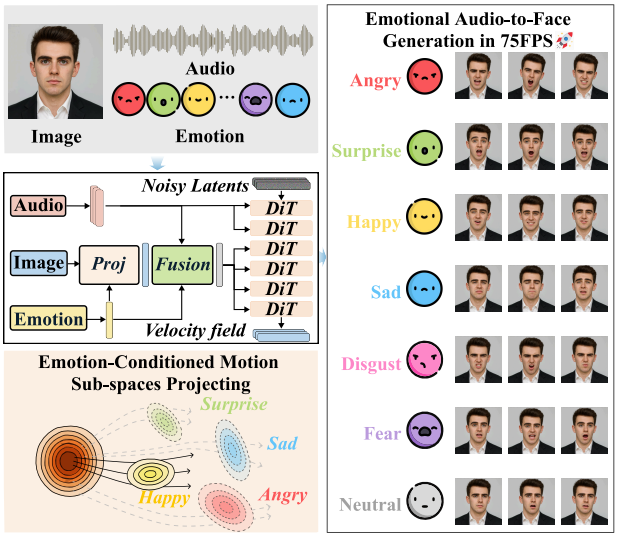


Figure 1: Our method synthesizes facial motion in real time while preserving lip-sync accuracy and emotional fidelity.

1 Introduction

Audio-driven talking-head synthesis (*Audio2Face*) has matured from a proof of concept into an enabling technology for virtual assistants, remote collaboration, and digital entertainment. Early systems focused almost exclusively on **lip-sync accuracy** and approached human performance through adversarial training and synchronization losses (Chen et al. 2019; Prajwal et al. 2020; Wang et al. 2021a; Ji et al. 2022; Zhang et al. 2023; Yu et al. 2023). As digital humans move toward emotionally rich, interactive applications, three additional requirements have become indispensable:

- **Emotional fidelity.** Facial behaviour must faithfully convey the speaker's emotion (e.g., down-turned lip corners for sadness, elevated zygomatic muscles for joy) instead of remaining expression-agnostic.
- **Multimodal coordination.** Audio, affect, and head-pose cues should be fused through a unified mechanism that captures high-order correlations, rather than being injected via separate task-specific branches.
- **Real-time performance.** Consumer devices allow only a few milliseconds per frame; multi-hundred-step diffusion samplers or large transformer stacks exceed this budget by one or two orders of magnitude.

To enable emotion control, many recent GAN-based methods explicitly factorize facial motion into lip articulation, expression, and head pose, driving the "expression branch" with either categorical labels or audio-derived prosody (Tan et al. 2024; Hong et al. 2025; Feng et al. 2024; Zhou et al. 2025). Although this design grants per-component editability, it tacitly assumes independence between articulation and affect. In reality, the two are strongly

*Corresponding author

coupled—sadness narrows the mouth aperture, while joy raises lip corners-so factorisation breaks this correlation and often yields videos that fail to convey the intended emotion.

Recent diffusion models learn the joint motion distribution and thereby capture articulation-affect dependencies (Xu et al. 2024; Cui et al. 2025b; Yang et al. 2025; Ji et al. 2025; Lin et al. 2025; Jiang et al. 2025; Chen et al. 2025; Shen et al. 2025). However, their remarkable fidelity comes with iterative samplings and heavy decoders. For example, Hallo (Xu et al. 2024) uses a 40-step diffusion process with a 2B parameter UNet, achieving 0.55 FPS inference speed on a single Ampere-based 80GB GPU. Moreover, most diffusion pipelines inject each modality via separate branches or naive feature concatenation. Every additional cue (e.g., eye gaze, speaking style) necessitates architectural surgery and often yields modality interference.

To overcome these challenges, we treat facial motion trajectories as points on a continuous latent space learned by an auto-encoder network. Motivated by the observation that articulation and emotion are intrinsically coupled, we posit that emotion-conditioned facial-motion latents concentrate in a compact sub-space of the motion latent space.

Specifically, we propose an emotion-conditioned Audio2Face framework, which is composed of three main modules: a motion dictionary with emotion-conditioned subspaces, an implicit multimodal fusion block, and a ten-step flow-matching sampler. Firstly, we learn a motion dictionary of orthogonal bases through a self-supervised autoencoder, then carve emotion-conditioned sub-spaces within the motion space. Secondly, we introduce a layer-progressive cross-attention multimodal fusion module that treats the current latent state as a query and the audio, emotion, and reference motion embeddings as keys/values. By selectively injecting cross-modal context in the deeper transformer layers while preserving low-level articulation in shallow layers, this module directly modulates the latent velocity field without bespoke branches. Thirdly, we train a velocity field predictor with a DiT architecture, and reduce generation to integrating a first-order ODE; only ten reverse steps and $\leq 180M$ parameters are sufficient to synthesize a complete motion-latent trajectory, achieving up to 75 FPS inference speed on a 3090.

To train the entire framework, we follow a three-stage curriculum. Stage I learns the motion dictionary on a 320K-clip corpus that merges VFHQ (Xie et al. 2022), HDTF (Zhang et al. 2021), and RAVDESS (Livingstone and Russo 2018). Multi-scale perceptual loss enforces global coherence, while mouth- and eye-centric patch discriminators with feature-matching objectives sharpen the regions most critical to human perception. Stage II fine-tunes on the emotion-annotated MEAD dataset (Wang et al. 2020) to shape the emotion-conditioned sub-spaces, using segment-level soft attention and mean-matching constraints to preserve lip-speech alignment and expression intensity; a soft-orthogonality regulariser prevents different emotion subspaces from collapsing. Stage III jointly trains the multimodal fusion module and the flow-matching transformer, so that under any combination of audio, emotion, and reference motion, the network predicts a denoising velocity field, completing the pathway from raw conditions to real-time, emotionally faithful talking heads.

We evaluate our method on MEAD (Wang et al. 2020) and RAVDESS (Livingstone and Russo 2018), comparing it against state-of-the-art GAN-based and diffusion-based baselines. Furthermore, we provide detailed analysis to validate the effectiveness of each component. We provide more video results regarding out-of-distribution generalization and emotion control in the supplementary video.

To summarize, our contributions are as follows:

- *Emotion-conditioned sub-spaces*. We introduce the first explicit learning of emotion-parameterised sub-spaces in the motion latent domain, preserving articulation-affect dependencies and improving naturalness.
- *Layer-progressive multimodal fusion*. A cross-attention fusion module injects audio and emotion into the latent velocity field via adaptive layer normalization and gating, enabling flexible control without architectural redesign.
- *Real-time flow-matching architecture*. Combining the motion dictionary with a ten-step sampler delivers the first Audio2Face system that simultaneously satisfies lip-sync accuracy, affective fidelity, and around 75 FPS inference speed, outperforming state-of-the-art baselines in both emotion accuracy, visual quality, and speed.

2 Related Work

Currently, audio-driven talking head generation methods can be divided into two categories: non-diffusion-based and diffusion-based.

2.1 Non-diffusion-based Audio-Driven Talking Head Generation

Early audio-driven methods focused on lip-sync accuracy, adopting end-to-end architectures with adversarial training and direct frame-wise generation via extra facial representations, e.g., facial landmarks (Chen et al. 2019; Zhou et al. 2020; Liu et al. 2023) and 3DMMs (Ren et al. 2021; Zhang et al. 2021, 2023). Recent works seek to learn a disentangled representation of facial motion, such as lip and non-lip (Yu et al. 2023), expression and head pose (Zhou et al. 2021), emotion and articulation (Tan et al. 2024; Feng et al. 2024; Wang et al. 2025a) to improve controllability. However, they ignore the intrinsic coupling between articulation and affect, leading to emotion-flat results. Our method learns a motion dictionary with emotion-conditioned sub-spaces, which preserves the natural articulation-affect coupling and improves expressiveness.

2.2 Diffusion-based Audio-Driven Talking Head Generation

Diffusion-based methods have recently emerged as a powerful framework for audio-driven talking head generation, learning the joint distribution of audio and facial motion (Tian et al. 2024; Cui et al. 2025a; Yang et al. 2025). These methods typically employ a diffusion model to learn the conditional distribution of facial motion given audio signals, such as in Hallo (Xu et al. 2024), Hallo3 (Cui et al.

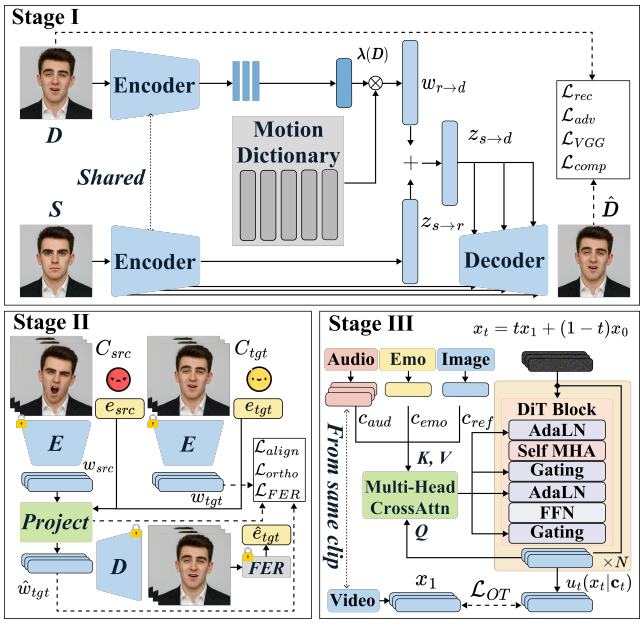


Figure 2: Method Overview. We follow a three-stage training strategy to progressively inject scale, affect, and multimodality. Stage I learns the motion dictionary $\mathbf{M}$ with an auto-encoder network. Stage II carves emotion-conditioned sub-spaces $\mathcal{S}_e$ inside the latent space. Stage III trains the multimodal fusion block and the flow-matching sampler to generate facial motion from audio and emotion conditions.

2025b), MegActor (Yang et al. 2025), and Sonic (Ji et al. 2025). They achieve high fidelity and naturalness by leveraging the diffusion process to generate realistic facial motion. However, these methods often require hundreds of diffusion steps and large UNet decoders, leading to slow inference speeds. Very recent works like Loopy (Jiang et al. 2025) and FLOAT (Ki, Min, and Chae 2024) have attempted to improve the efficiency of diffusion-based methods by utilizing a latent diffusion model (Rombach et al. 2022) or flow matching (Lipman et al. 2023). However, these methods typically inject audio and emotion through separate branches or concatenation, which can lead to interference between modalities and limit the expressiveness of the generated motion (Ki, Min, and Chae 2024). Our method addresses these limitations by introducing a unified multimodal fusion module that directly modulates the latent velocity field with audio and emotion signals, enabling flexible control without architectural redesign. Additionally, we employ a ten-step flow-matching sampler that achieves real-time performance while maintaining high fidelity and naturalness.

3 Method

Given a source image S, a target audio sequence $a^{1:T}$ and an emotion condition $e \in \mathbb{R}^7$ (probabilities of seven emotions: *angry*, *surprise*, *happy*, *sad*, *disgust*, *fear*, *neutral*), our goal is to synthesize a video $\hat{D}^{1:T}$ that synchronizes the lip motion with the audio and reflects the specified emotion. The overall architecture and training strategy are shown in Fig. 2.

3.1 Preliminaries

To learn a target distribution q over $\mathbb{R}^d$, *flow matching* (Lipman et al. 2023) constructs a probability path$(p_t)_{0\le t\le 1}$ from a simple distribution p to q by learning a time-dependent velocity field $u_t : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$, which defines an ODE:

$$\frac{\mathrm{d}}{\mathrm{d}t}\psi_t(x) = u_t(\psi_t(x)),\ \psi_0(x) = x \tag{1}$$

where $\psi_t(x)$ is the trajectory of a point x at time t. The goal is to learn a velocity field that approximates the denoising process of a diffusion model, such that the distribution of $\psi_1(x)$ matches the target distribution q. Therefore, the training objective is to minimize the following loss:

$$\mathcal{L}_{FM} = \mathbb{E}_{x\sim p_t, t\sim U(0,1)}\|u_t^\theta(x) - u_t(x)\|^2 \tag{2}$$

Flow matching adopts a *conditional optimal-transport path* and defines the random variable X_t at time t as $X_t = tX_1 + (1-t)X_0 \sim p_t$. Take $x = X_{t|1}$ and solve the ODE in Eq. 1 to obtain the *conditional velocity field* $u_t(x|x_1)$, and therefore the training objective in Eq. 2 can be rewritten as:

$$\mathcal{L}_{CFM} = \mathbb{E}_{X_0\sim\mathcal{N}(0,I), X_1\sim q, t\sim U(0,1)}\|u_t^\theta(X_t) - (X_1 - X_0)\|^2 \tag{3}$$

We note that it can be proved that $\mathcal{L}_{FM}$ and $\mathcal{L}_{CFM}$ are equivalent, sharing the same gradients to learn u_t^θ.

3.2 Motion Dictionary with Emotion-Conditioned Sub-Spaces

Unlike implicit-keypoints-based methods (Siarohin et al. 2019; Wang, Mallya, and Liu 2021) that utilize implicit keypoints to represent facial motion, we learn a motion latent space that captures the facial motion z parameterized by a *motion dictionary* $\mathbf{M}$. Specifically, given a facial image S, we obtain a latent code $z_s \in \mathbb{R}^d$ and decompose it into an identity component $z_{s\to r}$ and a facial motion latent $w_{r\to s}$, formulated as $z_s = z_{s\to r} + w_{r\to s}$. Following (Wang et al. 2024) we learn a motion dictionary $\mathbf{M} \in \mathbb{R}^{N\times d}$ that consists of a set of orthogonal bases $\{\mathbf{m}_i\}_{i=1}^N$ in the latent space. The motion trajectory $w_{r\to s}$ can be represented as a linear combination of these bases:

$$w_{r\to s} = \mathbf{M}^T \cdot \boldsymbol{\lambda}(S) = \sum_{i=1}^{N} \lambda_i(S)\mathbf{m}_i \tag{4}$$

Semantically, each base $\mathbf{m}_i$ represents a specific basic motion pattern, and the source-dependent coefficients $\lambda_i(S)$ control the contribution of each base to the overall motion trajectory. We note that the motion dictionary $\mathbf{M}$ is learned in a self-supervised manner, without requiring any explicit annotations, detailed in Sec. 3.4.

To incorporate emotion into the motion generation process, we carve emotion-conditioned sub-spaces $\mathcal{S}_e$ inside the dictionary. Specifically, a learnable selector $\Phi(e) \in \mathbb{R}^{k\times N}$ chooses k directions and yields the bases $\mathbf{B}_e \in \mathbb{R}^{k\times d}$ of the emotion-conditioned sub-space $\mathcal{S}_e$:

$$\mathbf{B}_e = \Phi(e)\cdot\mathbf{M},\ \mathbf{B}_e\mathbf{B}_e^T = I_k \tag{5}$$

spanning the sub-space $\mathcal{S}_e = \mathrm{span}(\mathbf{B}_e^T) \subset \mathbb{R}^d$. Intuitively, $\mathcal{S}_e$ captures motion patterns that co-vary with emotion e while inheriting orthogonality from $\mathbf{M}$.

With the learned emotion-conditioned sub-spaces, we can transform a motion trajectory $w_{r\to s}$ originating from emotion e_{src} to a target emotion e_{tgt} by projecting it into the corresponding sub-spaces:

$$\hat{w}_{r\to s} = (I - \mathbf{P}_{e_{src}})w_{r\to s} + \mathbf{P}_{e_{tgt}} w_{r\to s},\ \mathbf{P}_e = \mathbf{B}_e^T \mathbf{B}_e \quad (6)$$

where $\mathbf{P}_e$ is the projection matrix onto the sub-space $\mathcal{S}_e$. This operation preserves the identity-dependent component while seamlessly replacing the affective part of the motion trajectory with the target emotion, thus preserving the natural articulation-affect coupling.

3.3 Flow Matching in Motion Latent Space

To generate facial motion from the emotion-conditioned sub-spaces, we employ a flow matching algorithm (Lipman et al. 2023) that learns a velocity field in the motion latent space. The learned velocity field $u_t^\theta(x_t|\mathbf{c}_t)$ maps the noisy motion latent x_t and the condition signal $\mathbf{c}_t$ fused by audio and emotion conditions, to a target velocity field at sampled time $t \in [0, 1]$. With $u_t^\theta(x_t|\mathbf{c}_t)$, we can generate a motion trajectory conditioned with input audio signal $a^{1:T}$ and emotion condition e by solving the ODE in Eq. 1.

As Fig. 2 shows, we adopt a DiT-like transformer (Peebles and Xie 2023) to model the velocity field predictor, which consists of a series of transformer blocks that process the input motion latent x_t and the condition signal to produce the velocity field $u_t^\theta(x_t|\mathbf{c}_t)$.

At each time step t, the *condition signal* $\mathbf{c}_t = \mathcal{F}(c_{\text{aud}}, c_{\text{emo}}, c_{\text{ref}})$ is constructed by an implicit multimodal fusion module $\mathcal{F}$:

- c_{aud}—audio embedding from a frozen Wav2Vec2 encoder (Baevski et al. 2020).
- c_{emo}—predicted from audio via pre-trained predictors (Pepino, Riera, and Ferrer 2021) or explicit labels e.
- $c_{\text{ref}} = \hat{w}_{r\to s}$—reference motion projected into $\mathcal{S}_e$ to ensure emotional consistency.

Audio, reference-motion, and emotion conditions are first projected to a common channel space. For shallow layers($l < 2$), only audio conditions are involved to avoid early interference; deeper layers fuse all three modalities as keys/values and perform multi-head cross-attention. A depth-aware blend $\mathbf{c} = (1 - \lambda)c_{aud} + \lambda\, c_{\text{mix}}$ with $\lambda = (l - 2)/(L - 2)$ smoothly increases multimodal influence at the l-th layer. Then the fused context $\mathbf{c}$ is augmented with a sinusoidal time embedding and fed to a 6-D predictor that outputs LayerNorm scale γ, shift β, and residual gate g. These parameters modulate both the self-attention and feed-forward layers, enabling fine-grained control while keeping the residual path.

Formally, this process can be expressed as Algorithm 1 in the supplementary. Our novel design allows us to capture the high-order correlations between different modalities and enables flexible control over the generated motion trajectory.

3.4 Training Strategy

Our training curriculum comprises three stages that progressively introduce *scale*, *affect*, and *multimodality*.

Stage I: Motion Dictionary Learning In the first stage, we learn the motion dictionary $\mathbf{M}$ from a large-scale corpus that merges VFHQ (Xie et al. 2022), HDTF (Zhang et al. 2021), and RAVDESS (Livingstone and Russo 2018). We train the auto-encoder to reconstruct the motion latent $z_{s\to d}$ from the source image S and the target image D in a self-supervised manner. The training objective is to minimize the reconstruction loss, including an L_1 loss and a multi-scale perceptual loss (Johnson, Alahi, and Fei-Fei 2016) via a pre-trained VGG network (Simonyan and Zisserman 2014). Furthermore, to enhance the fidelity of the mouth and eye regions, we calculate facial component loss that focuses on the mouth and eye regions, respectively, following (Wang et al. 2021b). The overall loss function for this stage is:

$$\begin{aligned}\mathcal{L}_{stage1} =& \mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{VGG}\mathcal{L}_{VGG} \\ &+ \lambda_{mouth}\mathcal{L}_{mouth-comp} + \lambda_{eye}\mathcal{L}_{eye-comp}\end{aligned} \quad (7)$$

Stage II: Emotion-Conditioned Sub-Space Learning In the second stage, we build emotion-conditioned sub-spaces $\mathcal{S}_e$ within the motion dictionary $\mathbf{M}$ using the emotion-annotated MEAD dataset (Wang et al. 2020). In this stage, we freeze the motion dictionary $\mathbf{M}$ and the auto encoder learned in Stage I, and only train the sub-spaces projector Φ. We randomly sample pairs of clips with the same speaker and content but different emotions from the MEAD dataset. We denote the source emotion as e_{src}, the corresponding clip as C_{src}, and the target emotion as e_{tgt}, with the corresponding clip as C_{tgt}. We first extract the motion latent $w_{src} \in \mathbb{R}^{T\times d}$ from the source clip C_{src}, and project it to the target emotion sub-space $\mathcal{S}_{e_{tgt}}$ from $\mathcal{S}_{e_{src}}$ to obtain the reference motion $\hat{w}_{tgt}$, following Eq. 6. Due to the lack of frame-wise alignment between the source and target clips, we employ a segment-level soft attention mechanism to enable supervision with the target clip C_{tgt}. Specifically, we compute the segment-level soft attention matrix A, which measures the similarity between the source and target segments, and use it to compute the aligned motion latent sequence $\hat{w}_{tgt}^{aligned}$. With the aligned motion latent sequence, we can compute the soft alignment loss $\mathcal{L}_{align}$ to align the mean of the source and target motion latents, formulated as:

$$A = \text{softmax}\left(w_{src}\hat{w}_{tgt}^T/\sqrt{d}\right) \quad (8)$$

$$\tilde{w}_{tgt} = A \cdot \hat{w}_{tgt} \quad (9)$$

$$\mathcal{L}_{align} = \|\frac{1}{T}\sum_{t=1}^{T}\tilde{w}_{tgt}(t) - \frac{1}{T}\sum_{t=1}^{T} w_{tgt}(t)\|_2^2 \quad (10)$$

To ensure that the emotion-conditioned sub-spaces $\mathcal{S}_e$ are orthogonal to each other, we introduce a soft-orthogonality regularizer $\mathcal{L}_{ortho}$ that encourages the bases of different emotion sub-spaces to be orthogonal. We note that we do not restrict each emotion sub-space to be orthogonal to each other, as this would limit the expressiveness of the model. In practice, a high correlation between different emotion sub-spaces is often observed, which can be beneficial for the model to learn the articulation-affect coupling.

$$\mathcal{L}_{ortho} = \sum_{e_1,e_2\in E, e_1\neq e_2} \|\mathbf{B}_{e_1}^T \mathbf{B}_{e_2}\|_F^2 \quad (11)$$

Method	FPS	MEAD						RAVDESS					
		PSNR	FID	FVD	LSE-C	M/F-LMD	Acc_{emo}	PSNR	FID	FVD	LSE-C	M/F-LMD	Acc_{emo}
Hallo	0.55	21.87	20.69	85.11	6.58	2.48/2.52	63.89	16.97	18.13	181.52	5.72	1.89/2.87	41.02
EchoMimic	0.76	21.69	20.56	142.47	5.76	2.43/2.59	75.00	15.88	17.94	299.01	4.49	1.90/2.88	42.50
Sonic	1.82	21.33	22.62	56.59	**8.09**	2.67/2.42	77.78	16.94	17.65	91.13	5.87	1.82/2.82	43.59
FLOAT	41.37	20.45	22.10	75.93	6.93	2.74/3.14	64.10	15.21	17.93	85.66	5.37	1.84/3.53	45.00
SadTalker	9.31	21.95	44.15	81.10	7.66	2.62/2.70	70.00	16.49	23.85	141.10	**6.04**	2.09/3.15	42.50
EAT	24.14	18.00	33.29	232.25	7.76	2.52/3.50	75.00	17.24	36.15	228.33	5.65	2.03/2.70	25.00
EDTalk	8.03	17.58	33.67	294.48	7.95	2.99/4.34	69.44	15.21	52.35	219.53	6.03	2.09/3.67	35.00
Ours	**91.23**	**22.58**	**19.96**	**50.39**	7.92	**2.21/2.34**	**82.05**	**17.52**	**15.04**	**62.15**	5.91	**1.66/2.40**	**62.50**
GT	-	-	-	-	7.98	-	95.00	-	-	-	5.93	-	82.50

Table 1: Quantitative results on MEAD and RAVDESS datasets. For FPS, PSNR, LSE-C, and Acc_{emo}, higher is better; for FID, FVD, M/F-LMD, lower is better. The best results are in **bold**. FPS is tested with a single Ampere-based 80GB GPU.

Additionally, we introduce an emotion recognition loss $\mathcal{L}_{FER}$ calculated with (Toisoul et al. 2021):

$$\mathcal{L}_{FER} = \text{CrossEntropy}(\text{FER}(\text{DEC}(\hat{w}_{tgt})), e_{tgt}) \quad (12)$$

The overall loss function for this stage is formulated as:

$$\mathcal{L}_{stage2} = \lambda_{align}\mathcal{L}_{align} + \lambda_{ortho}\mathcal{L}_{ortho} + \lambda_{FER}\mathcal{L}_{FER} \quad (13)$$

Stage III: Multimodal Fusion and Flow Matching In the third stage, we freeze both $\mathbf{M}$ and $\mathcal{S}_e$, jointly train the multimodal fusion module and the flow matching transformer to predict the denoising velocity field $u_t^\theta(x_t|\mathbf{c}_t)$ with all training datasets. We sample fixed-length windows-L frames from the training set as well as the corresponding audio signals, and extract the emotion signals via a pre-trained emotion predictor (Pepino, Riera, and Ferrer 2021). To enable chunk-by-chunk inference for streaming, we involve a sliding window strategy that feeds the previous L' frames and audio clips together with the current frame to the model. Following the derivation in Sec. 3.3, we optimize the conditional flow matching loss $\mathcal{L}_{CFM}$ in Eq. 3:

$$\mathcal{L}_{stage3} = \|u_t^\theta(x_t|\mathbf{c}_t) - (x_1 - x_0)\|_1 \quad (14)$$

where $x_t = tx_1 + (1-t)x_0$ is the noised motion latent at time t, and $x_0 \sim N(0, I)$ denotes the noise sampled from a Gaussian distribution, and x_1 is the target motion latent sampled from the emotion-conditioned sub-space $\mathcal{S}_e$. We sample the time step from a logit-norm distribution $t \sim \text{LogitNorm}(0, 1)$ instead of a uniform distribution, as (Wang et al. 2025b) point out that sampling more frequently in the middle time steps can improve the generation quality of the model. Furthermore, to enable classifier-free guidance, we randomly dropout the condition signal $c_{aud}, c_{emo}, c_{ref}$ with a probability of 0.1 during training.

$$\begin{aligned}\tilde{u}_t =&(1+\gamma_a+\gamma_e)u_t^\theta(x_0, \mathbf{c}_t) \\ &-\gamma_a u_t^\theta(x_0, \mathbf{c}_t|c_{aud}=0) - \gamma_e u_t^\theta(x_0, \mathbf{c}_t|c_{emo}=0)\end{aligned} \quad (15)$$

Guidance scales γ_a, γ_e can be applied to the condition signal during inference following (Dao et al. 2023) as Eq. 15.

4 Experiments

Baselines We select several state-of-the-art methods for comparison, including GAN-based methods: **SadTalker** (Zhang et al. 2023), **EAT** (Gan et al. 2023), **EDTalk** (Tan et al. 2024) and Diffusion-based methods: **Hallo** (Xu et al. 2024), **EchoMimic** (Chen et al. 2025), **Sonic** (Ji et al. 2025) and **FLOAT** (Ki, Min, and Chae 2024). We fine-tuned these methods on the same training set as ours to ensure a fair comparison. For FLOAT, we set $\gamma_e = 2$ for emotional talking head generation as suggested in the original paper.

Metrics We evaluate the generated videos in terms of emotion accuracy, lip-sync accuracy, video quality, and inference speed. SadTalker, EAT, and EDTalk are evaluated in 256×256 resolution, while others are evaluated in 512×512.

- For emotion accuracy, we use a pre-trained Emotion-Fan (Meng et al. 2019) to predict the emotion of generated videos, and compute the accuracy Acc_{emo}. Following EAT (Gan et al. 2023), we fine-tune the Emotion-Fan model on MEAD and RAVDESS, respectively.
- For lip-sync accuracy, we calculate lip-sync error confidence (*LSE-C*) using a pre-trained SyncNet (Chung and Zisserman 2016). In addition, we compute the distance between the landmarks of the mouth (*M-LMD*) and the whole face (*F-LMD*) to evaluate the lip-sync accuracy and the pose and expression accuracy of the generated videos, respectively.
- For visual quality, we measure *PSNR* and *FID* (Heusel et al. 2017). Additionally, we compute 16-frame Fréchet Video Distance (*FVD*) (Unterthiner et al. 2018) to evaluate the temporal consistency of the generated videos.
- For inference speed, we measure average frames-per-second (*FPS*) to generate 10-second video clips with a single Ampere-based 80GB GPU.

4.1 Emotional Talking Head Generation

We conduct experiments on MEAD (Wang et al. 2020) and RAVDESS (Livingstone and Russo 2018) to evaluate the ability to generate emotional talking heads. Each dataset is

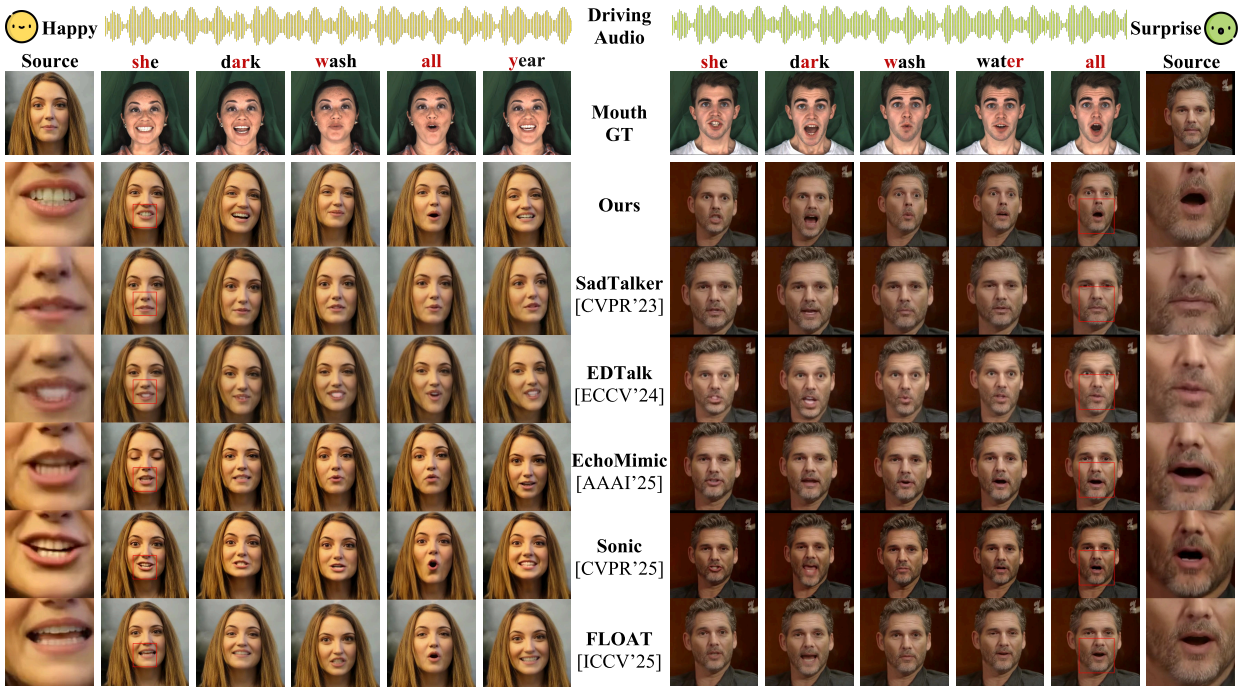


Figure 3: Qualitative results of emotional talking head generation. The first row shows the input audio and the target emotion label; the second row visualizes reference ground-truth mouth motion. Subsequent rows present outputs from our method and recent baselines. The first and last column provides zoom-ins of the mouth region for perceptual detail.

split *8:1:1* into train/val/test, ensuring the test set contains speakers unseen during training. Quantitative and qualitative results are shown in Tab. 1 and Fig. 3, respectively.

GAN-based methods disentangle lip, expression, and pose streams and drive each with separate losses, enabling local editing but *breaks articulation-affect coupling*. These methods struggle to generate realistic emotional talking heads, as reported in terms of PSNR and FVD. Diffusion-based methods learn the joint motion distribution and yield impressive realism, yet depend on hundreds of sampling steps and large style-U-Nets. Additionally, they fail to express emotion accurately, as shown in Tab. 1, achieving lower than 50% emotion accuracy on RAVDESS.

Our approach attains the best visual quality on both datasets. Specifically, on MEAD we reach **22.58 dB** PSNR and **19.96/50.39** FID/FVD, surpassing the baselines by large margins; similar gains hold on RAVDESS. Compared with diffusion-based methods, our method achieves a 166× speedup compared to Hallo and a 50× speedup over Sonic, while maintaining high fidelity and naturalness.

In terms of lip-sync accuracy, our method achieves the lowest M/F-LMD on both datasets. Our method achieves a lower LSE-C than some methods, due to the richer head motion in the generated videos, leading to less accurate mouth feature extraction by SyncNet. However, our M-LMD is still the best among all methods, and the LSE-C is close to that of ground-truth videos, indicating a good lip-sync accuracy. Visual comparison in Fig. 3 and supplementary videos further confirms that our method can generate realistic emotional talking heads with natural lip-sync and head motion.

Method	Visual ↑	Lip-sync ↑	Acc_{emo} ↑
EAT	2.75 ± 0.17	2.98 ± 0.18	$79.00\% \pm 1.48\%$
EDTalk	1.98 ± 0.23	2.91 ± 0.16	$69.00\% \pm 2.20\%$
EchoMimic	3.76 ± 0.12	3.53 ± 0.16	$47.00\% \pm 1.65\%$
Sonic	4.30 ± 0.15	4.62 ± 0.11	$61.00\% \pm 1.48\%$
FLOAT	4.31 ± 0.16	4.44 ± 0.13	$56.00\% \pm 3.43\%$
Ours	$\mathbf{4.92 \pm 0.15}$	$\mathbf{4.86 \pm 0.14}$	$\mathbf{92.50\% \pm 0.97\%}$

Table 2: User study results on visual realism, lip-sync accuracy, and emotional accuracy.

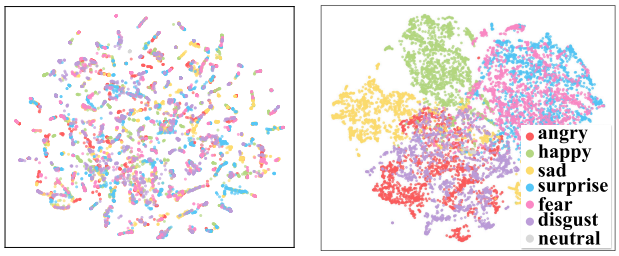
In terms of emotion accuracy, our method achieves a significantly higher emotion accuracy (**82.05%** on MEAD and **62.50%** on RAVDESS) than all other methods. This demonstrates that our method can effectively capture the articulation-affect coupling and generate high-quality emotional talking heads, thanks to the novel design of the emotion-conditioned sub-spaces and the layer-wise multi-modal fusion module. Qualitative results in Fig. 3 illustrate that our method can generate expressive emotional talking heads that accurately reflect the target emotion.

4.2 User Study

To further evaluate the quality of the generated videos, we conduct a user study on 50 participants. Each participant is shown 10 sequences of videos generated by our method

	PSNR	FID/FVD	LSE-C	M-LMD	Acc_{emo}
concat	18.90	22.75/97.59	6.93	2.93	33.33
self-attn	19.22	22.85/64.30	6.81	2.35	72.56
cross-attn	21.60	22.68/56.73	4.34	2.73	69.44
w/o $\mathcal{S}_e$	21.90	22.15/63.43	6.81	2.39	64.10
random $\mathcal{S}_e$	19.31	20.85/84.30	5.67	2.99	45.00
Ours	**22.58**	**19.96/50.39**	**7.52**	**2.21**	**82.05**

Table 3: Ablation study of layer-progressive cross-modal fusion module and emotion-conditioned sub-spaces.



(a) Original latent space (b) Emotion sub-spaces

Figure 4: t-SNE(d=2) visualization of the motion latent space. (a) shows the original motion latent space, and (b) shows the emotion-conditioned sub-spaces learned by our method. The different colors represent different emotions.

and several baselines (EAT, EDTalk, EchoMimic, Sonic, and FLOAT), sampled from the test dataset, and asked to score from 1 (worst) to 5 (best) in terms of visual realism and lip-sync accuracy and evaluate the emotional accuracy. The results are shown in Tab. 2. Our method outperforms the baseline methods in all three aspects, demonstrating its superiority in generating realistic emotional talking heads.

4.3 Ablation Study

Emotion-conditioned sub-spaces To verify that the emotion-conditioned sub-spaces $\mathcal{S}_e$ are effective in capturing the articulation-affect coupling, we compare our method with two variants: *w/o* $\mathcal{S}_e$, which directly uses the motion dictionary $\mathbf{M}$ without any emotion conditioning; and *random* $\mathcal{S}_e$, which selects random bases from $\mathbf{M}$. Table 3 shows that removing $\mathcal{S}_e$ drops emotion accuracy by $18\,\mathrm{pp}$, while random bases hurt every metric, failing to capture the emotion information and leading to poor lip-sync and emotion accuracy. This proves that $\mathcal{S}_e$ is essential for generating high-quality emotional talking heads, providing more than mere dimensionality reduction.

Figure 4 visualises the latent space with t-SNE (Maaten and Hinton 2008). In the original space (Fig. 4(a)), emotion clusters overlap heavily, indicating that the original representation is dominated by phonetic and speaker-specific factors rather than affective cues. In contrast, Fig. 4(b) shows the same samples after being projected to our learned emotion sub-spaces. The points now aggregate into seven distinguishable clusters. We notice that two pairs of emotion — angry vs. disgusted and fear vs. surprise — still exhibit partial overlaps. This is because t-SNE compresses a 512-D latent space into only two dimensions, inevitably discarding some discriminative variance. Furthermore, these emotion pairs share similar facial action patterns (e.g., brow-lowering for anger and disgust, and eye-widening for fear and surprise), which makes them hard to separate.

Cross-modal fusion module To validate the effectiveness of the cross-modal fusion module, we conduct an ablation study by comparing the performance of our method with and without the cross-modal fusion module. We evaluate several strategies: directly concatenate the audio and emotion conditions without any fusion mechanism (*concat*); apply self-attention on different modalities and fuse them with a linear layer (*self-attn*); and activate cross-attention across all layers of the DiT block (*cross-attn*). As Table 3 indicates, naive concatenation fails to model high-order interactions (emotion 33%), while self-attention helps lip-sync (better M-LMD) but not realism (FID stagnates). The *cross-attn* strategy improves the performance significantly, but still lags behind our method, especially in terms of emotion accuracy and lip-sync accuracy, confirming that early fusion disturbs phoneme timing. Our layer-progressive cross-attention scheme preserves temporal structure at lower layers and injects affective context higher up, yielding the best overall balance—-6.34 FVD, -0.14 M-LMD, and $+9.49\,\mathrm{pp}$ emotion accuracy over the strongest baseline. This design enables the model to better capture high-order correlations between audio and emotion signals while preserving the temporal structure of the motion latent, therefore generating emotionally expressive and realistic talking head videos.

5 Conclusion

We introduced an Audio2Face framework that jointly learns a motion dictionary and *emotion-conditioned sub-spaces*, captures articulation-affect coupling, and deploys a ten-step flow-matching sampler for real-time generation. Extensive experiments demonstrate that our method achieves state-of-the-art performance in terms of emotion accuracy, lip-sync accuracy, video quality, and inference speed, while ablations confirm the benefits of both the sub-space design and our layer-progressive fusion block. Our method produces natural, expressive talking heads at 75 FPS on a consumer-grade device. Beyond outperforming existing baselines, our design offers a scalable blueprint for multimodal character animation: new cues can be injected via the same fusion mechanism, while additional affective styles map to distinct sub-spaces without architectural surgery.

We also acknowledge the limitations of our current model, which is trained solely on English speech and a limited set of seven basic emotions. Furthermore, most training sequences involve near-frontal facial poses, limiting the model's generalization to extreme head rotations. As future work, we plan to expand the linguistic and emotional coverage of the dataset, incorporate pose-diverse training samples to enhance robustness to head movement, and explore text-conditioned generation to enable fully scriptable, multilingual character synthesis.

Acknowledgements

This work is supported by the National Key R&D Program of China under Grant No.2024QY1400, and the National Natural Science Foundation of China No. 62425604, 62502256. This work is also supported by Beijing Natural Science Foundation (L257006) and Tsinghua University Initiative Scientific Research Program & the Institute for Guo Qiang at Tsinghua University.

References

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33: 12449–12460.

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

Chen, Z.; Cao, J.; Chen, Z.; Li, Y.; and Ma, C. 2025. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2403–2410.

Chung, J. S.; and Zisserman, A. 2016. Out of time: automated lip sync in the wild. In *Asian Conference on Computer Vision*, 251–263. Springer.

Cui, J.; Li, H.; Yao, Y.; Zhu, H.; Shang, H.; Cheng, K.; Zhou, H.; Zhu, S.; and Wang, J. 2025a. Hallo2: Long-Duration and High-Resolution Audio-Driven Portrait Image Animation. In *International Conference on Learning Representations*.

Cui, J.; Li, H.; Zhan, Y.; Shang, H.; Cheng, K.; Ma, Y.; Mu, S.; Zhou, H.; Wang, J.; and Zhu, S. 2025b. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21086–21095.

Dao, Q.; Phung, H.; Nguyen, B.; and Tran, A. 2023. Flow matching in latent space. arXiv:2307.08698.

Feng, G.; Qian, Z.; Li, Y.; Jin, S.; Miao, Q.; and Pun, C.-M. 2024. Les-talker: Fine-grained emotion editing for talking head generation in linear emotion space. arXiv:2411.09268.

Gan, Y.; Yang, Z.; Yue, X.; Sun, L.; and Yang, Y. 2023. Efficient emotional adaptation for audio-driven talking-head generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22634–22645.

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. *Advances in Neural Information Processing Systems*, 30.

Hong, F.-T.; Xu, Z.; Zhou, Z.; Zhou, J.; Li, X.; Lin, Q.; Lu, Q.; and Xu, D. 2025. Audio-visual controlled video diffusion with masked selective state spaces modeling for natural talking head generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Ji, X.; Hu, X.; Xu, Z.; Zhu, J.; Lin, C.; He, Q.; Zhang, J.; Luo, D.; Chen, Y.; Lin, Q.; et al. 2025. Sonic: Shifting focus to global audio perception in portrait animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 193–203.

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*. Association for Computing Machinery.

Jiang, J.; Liang, C.; Yang, J.; Lin, G.; Zhong, T.; and Zheng, Y. 2025. Loopy: Taming Audio-Driven Portrait Avatar with Long-Term Motion Dependency. In *International Conference on Learning Representations*.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 694–711. Springer.

Ki, T.; Min, D.; and Chae, G. 2024. Float: Generative motion latent flow matching for audio-driven talking portrait. arXiv:2412.01064.

Lin, G.; Jiang, J.; Yang, J.; Zheng, Z.; and Liang, C. 2025. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2023. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*.

Liu, Y.; Lin, L.; Yu, F.; Zhou, C.; and Li, Y. 2023. Moda: Mapping-once audio-driven portrait animation with dual attentions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23020–23029.

Livingstone, S. R.; and Russo, F. A. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one*, 13(5): e0196391.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov): 2579–2605.

Meng, D.; Peng, X.; Wang, K.; and Qiao, Y. 2019. Frame attention networks for facial expression recognition in videos. In *2019 IEEE International Conference on Image Processing (ICIP)*, 3866–3870. IEEE.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Pepino, L.; Riera, P.; and Ferrer, L. 2021. Emotion Recognition from Speech Using wav2vec 2.0 Embeddings. In *Proc. Interspeech 2021*, 3400–3404.

Prajwal, K.; 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. In *Proceedings of the 28th ACM International Conference on Multimedia*, 484–492.

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/CVF International Conference on Computer Vision*, 13759–13768.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Shen, F.; Wang, C.; Gao, J.; Guo, Q.; Dang, J.; Tang, J.; and Chua, T.-S. 2025. Long-Term TalkingFace Generation via Motion-Prior Conditional Diffusion Model. In *Forty-second International Conference on Machine Learning*.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019. First order motion model for image animation. In *Advances in Neural Information Processing Systems*, volume 32, 7137–7147.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

Tan, S.; Ji, B.; Bi, M.; and Pan, Y. 2024. Edtalk: Efficient disentanglement for emotional talking head synthesis. In *European Conference on Computer Vision*, 398–416. Springer.

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. In *European Conference on Computer Vision*, 244–260. Springer.

Toisoul, A.; Kossaifi, J.; Bulat, A.; Tzimiropoulos, G.; and Pantic, M. 2021. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3(1): 42–50.

Unterthiner, T.; Van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; and Gelly, S. 2018. Towards accurate generative models of video: A new metric & challenges. arXiv:1812.01717.

Wang, B.; Zhu, X.; Shen, F.; Xu, H.; and Lei, Z. 2025a. PC-Talk: Precise Facial Animation Control for Audio-Driven Talking Face Generation. arXiv:2503.14295.

Wang, K.; Wu, Q.; Song, L.; Yang, Z.; Wu, W.; Qian, C.; He, R.; Qiao, Y.; and Loy, C. C. 2020. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, 700–717. Springer.

Wang, S.; Li, L.; Ding, Y.; Fan, C.; and Yu, X. 2021a. Audio2Head: Audio-driven One-shot Talking-head Generation with Natural Head Motion. In *IJCAI International Joint Conference on Artificial Intelligence*, 1098–1105.

Wang, T.-C.; Mallya, A.; and Liu, M.-Y. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10039–10049.

Wang, X.; Li, Y.; Zhang, H.; and Shan, Y. 2021b. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9168–9178.

Wang, Y.; Yang, D.; Bremond, F.; and Dantcheva, A. 2024. Lia: Latent image animator. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 10829–10844.

Wang, Z.; Zhang, P.; Qi, J.; Wang, G.; Ji, C.; Xu, S.; Zhang, B.; and Bo, L. 2025b. OmniTalker: One-shot Real-time Text-Driven Talking Audio-Video Generation With Multimodal Style Mimicking. arXiv:2504.02433.

Xie, L.; Wang, X.; Zhang, H.; Dong, C.; and Shan, Y. 2022. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 657–666.

Xu, M.; Li, H.; Su, Q.; Shang, H.; Zhang, L.; Liu, C.; Wang, J.; Yao, Y.; and Zhu, S. 2024. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. arXiv:2406.08801.

Yang, S.; Li, H.; Wu, J.; Jing, M.; Li, L.; Ji, R.; Liang, J.; Fan, H.; and Wang, J. 2025. Megactor-sigma: Unlocking flexible mixed-modal control in portrait animation with diffusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9256–9264.

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/CVF International Conference on Computer Vision*, 7645–7655.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3661–3670.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4176–4186.

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

Zhou, Z.; Quan, W.; Shi, H.; Li, W.; Wang, L.; and Yan, D.-M. 2025. GoHD: Gaze-oriented and Highly Disentangled Portrait Animation with Rhythmic Poses and Realistic Expressions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10914–10922.