Talking Head Generation with Probabilistic Audio-to-Visual Diffusion Priors

Zhentao Yu[1*] Zixin Yin[1,2*†] Deyu Zhou[1,3*†] Duomin Wang[1]
Finn Wong[1] Baoyuan Wang[1‡]
[1]Xiaobing.AI.
[2]The Hong Kong University of Science and Technology.
[3]The Hong Kong University of Science and Technology (Guangzhou).
(yuzhentao,wangduomin,wangwenlan,wangbaoyuan)@xiaobing.ai
zyinaf@connect.ust.hk, dzhou861@connect.hkust-gz.edu.cn

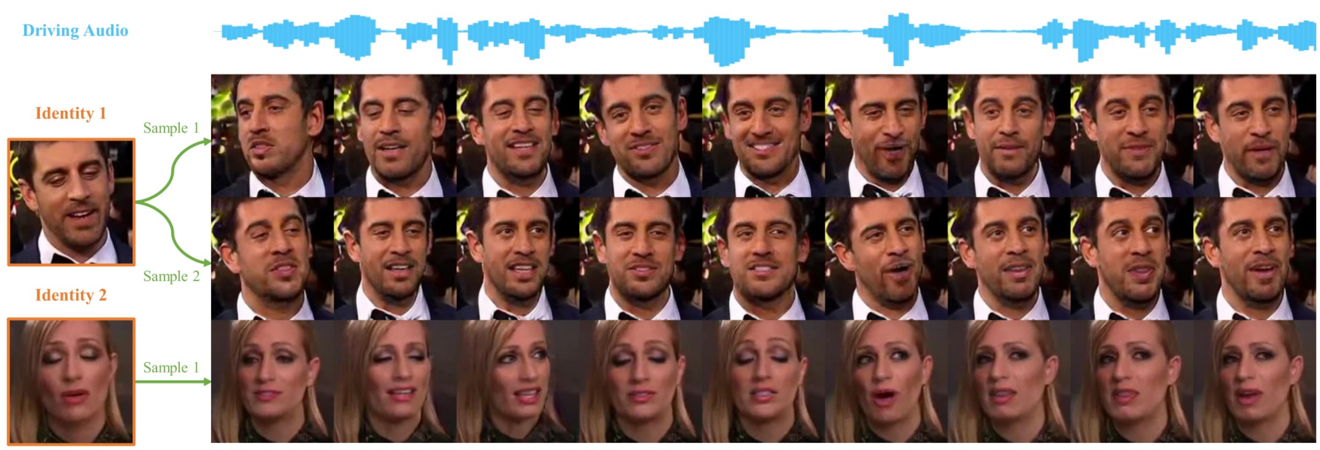

Figure 1. Given only an audio source and an arbitrary identity image, our system can generate a video with natural-looking and diverse facial motions (pose, expression, blink & gaze), while maintaining accurate audio-lip synchronization. Here we show randomly sampled sequences from our diffusion prior for two identities, note that the lip-irrelevant facial motion varies but the lip is still in-sync.

Abstract

We introduce a novel framework for one-shot audio-driven talking head generation. Unlike prior works that require additional driving sources for controlled synthesis in a deterministic manner, we instead sample all holistic lip-irrelevant facial motions (i.e. pose, expression, blink, gaze, etc.) to semantically match the input audio while still maintaining both the photo-realism of audio-lip synchronization and overall naturalness. This is achieved by our newly proposed audio-to-visual diffusion prior trained on top of the mapping between audio and non-lip representations. Thanks to the probabilistic nature of the diffusion prior, one big advantage of our framework is it can synthesize diverse facial motion sequences given the same audio clip, which is quite user-friendly for many real applications. Through comprehensive evaluations of public benchmarks, we conclude that (1) our diffusion prior outperforms auto-regressive prior significantly on all the concerned metrics; (2) our overall system is competitive with prior works in terms of audio-lip synchronization but can effectively sample rich and natural-looking lip-irrelevant facial motions while still semantically harmonized with the audio input.

1. Introduction

Audio-driven face reenactment and talking head generation have received raising attention due to the broad killer applications in movie production, gaming, virtual digital avatars, and potentially more while we move toward the era

*These authors have contributed equally to this work.
†This work was done when Zixin Yin and Deyu Zhou were interns at XiaoBing.AI.
‡Corresponding author.

of the metaverse. The past literature tries to advance this area from various perspectives. One line of research focuses on improving the generation quality originating from regular GAN-based methods [64, 65] to pre-trained StyleGAN [3] and to the most recent Neural Radiance Field (NeRF) based methods [13, 42]. Orthogonal to direction, another line of works emphasizes the importance of disentangled representation from the audio [32, 21, 66] for controlled generations. *i.e.*, [21] can predict the emotion from the audio input while [66] can decouple the speech signal into speaker identity and the phonetic content. A closely related set of works targeting more granular controlled talking head generation by providing additional input signals. For example, PC-AVS [65] requires a separate video to provide the pose signal in addition to the audio clip but does not support the control of expression and blinking. To remedy this, GC-AVT [27] requires inputting additional driving video for expression in addition to pose and audio driving clip. The technical challenge behind those approaches is how to faithfully transfer the desired driving signal (*i.e.*, pose, expression, audio) into the results without affecting each other through intrinsic disentanglement. Although encouraging results have been reported, we argue such a setting is **not practical** for broader applications. It is quite challenging, or at least labor-intensive, for novice users to find the "best" individual driving sources for each controlled dimension to make the final talking head video overall look not only lifelike but also coherent from semantic and emotional perspectives. Therefore, it is both more practical and generalizable if the setting only requires an audio signal as the driving source and expects a data-driven model to sample reasonable other facial motions irrelevant to lips.

By no means, we are the first to advocate an audio-only driving setup for talking head generation. Audio2head [54] predicts the poses from audio input with an LSTM network. [29] employs an autoregressive model by assuming the poses are jointly determined by the audio and past head motions along the sequence, although the results are encouraging, the model can't generalize due to their person-specific training strategy. FACIAL [61] could infer the blink, but requires a reference video input rather than a one-shot reference image. There are other related works along this direction, but they either only infer one facial attribute (*i.e.*, emotion in EVP[21], pose from [54]) using ad-hoc methods, or they simply do not support inferring other facial motions [38], or treat it as a block-box mapping model [3] without explicitly respecting the one-to-many mapping nature between audio and other visual facial attributes (pose, expression, blink). A more principled solution is desired to consolidate this line of work.

In this paper, we introduce a novel framework that can holistically infer all the non-lip-related facial attributes from the audio input while maintaining accurate synchronization between audio and the corresponding visual lip motions. This is achieved by two important learning steps in our pipeline. (1) A pre-trained identity-irrelevant facial motion representation can help decouple the lip and non-lip representations. To promote this learning, we employ a novel orthogonal loss on top of the modified facial reenactment framework [6]. The disentanglement enables to generate one-to-one mapping with lip representations to ensure the synchronization, and generate richer non-lip representations with a one-to-many mapping. (2) A novel audio-to-visual diffusion prior model is introduced to address the probabilistic sampling from audio representations to the above-learned non-lip representation. This prior is expected to solve the one-to-many mapping and provide diverse results during the inference stage. The entire pipeline can be easily built up on top of the existing framework, such as PC-AVS[65] without heavily retraining every component. To sum up, we make the following contributions:

- To our best knowledge, we are the first to holistically predict non-lip facial motions based on audio input only, providing good usability for reenactment or dubbing applications without extra driving video sources. Our method addresses the intrinsic one-to-many challenge in a probabilistic way, allowing diverse and realistic facial motion generation under the same audio input, as shown in Fig. 1.
- We leverage the pre-trained visual identity-irrelevant facial motion representations. And further, learn disentangled lip-related and lip-irrelevant representations through a novel orthogonal loss on top of PC-AVS[65]. A powerful diffusion prior model is then introduced to effectively infer all lip-irrelevant facial motions for a given audio segment in the representation space.
- We systematically evaluate the naturalness and diversity of the results with new metrics, which paves a way for future studies. Meanwhile, results show that our method can produce natural-looking head poses and facial motions without hurting the audio-lip synchronization. Our model will be released.

2. Related Works

Audio-Visual Cross-modal Learning Cross-modal representation learning is a long-standing research topic, ranging from speech enhancement [11], speech source separation [26] to synchronization [8, 22] and other speech disentanglement [34, 21, 32]. Among them, EVP [21] used cross-modal supervision to disentangle speech content and emotion from the audio signal with landmark as the intermediate representation. Recently, CMC [49] discussed how multi-view "modality" can be jointly unitized to boost intrinsic representation through contrastive learning, rather than predictive (or reconstruction) learning, it also demonstrated that the more views, the better. Concurrently, MMV [2] introduced different modality embedding graphs for effective cross-modal representations, again through contrastive learning. More recently, HCMoCo [16] extended similar ideas with a hierarchical strategy to learn different levels of representations for human-centric perception tasks. Although impressive results were reported, it is still unclear how it performs on face analysis tasks, not mention to on synthesis tasks. Our work shares a similar spirit but a different purpose, where we first pre-train a non-identity visual representation which then helps learn a lip and non-lip space, the latter further serves as the upper-bound learning target for subsequent audio-to-visual diffusion prior.

Face Reenactment & Talking Head Generation Face Reenactment is designed to transfer part or full facial motion from a driving source to the target video with good ID-preserved appearance and background. It can be further divided into two categories depending on whether the driving source is from video [17, 18, 5, 48, 24, 36, 59, 19, 60, 6, 56] or audio [52, 47, 62, 21, 55, 37, 63, 30]. Among them, audio-driven face reenactment generally aims to edit the mouth regions of the target video in order to match the input audio while leaving other facial attributes mostly unchanged, *i.e*., pose. EVP [21] tries to infer the non-rigid facial expression in addition to lip motion from the audio input. A closely related line of work is the audio-driven talking-head generations [38, 66, 61, 3] where only one target reference face is given, hence, other face attributes including pose, expression, blink and gaze have to be either explicitly given [65, 27] or partly inferred through statistical methods [66, 31, 61, 29, 3, 47]. Specifically, Lu *et al*. [29] employed an auto-regressive model while Min *et al*. [31] leveraged normalized flow prior to predicting a natural-looking pose sequence from the input audio, both showed encouraging results. Compared with them, we aim to infer more diverse facial motions including pose, expression, and even blink and gaze in a holistic manner through an audio-to-visual diffusion prior model.

Diffusion Generative Models The diffusion model [44, 15, 45], which is a likelihood-based model consisting of cascading denoising autoencoders, has recently shown great success in numerous generative tasks with different modalities including image [10, 35, 40, 41], audio [25], video [43], and motion [46]. To name a few, DDPM [15] explored the diffusion model for unconditional image generation. GLIDE [35] introduced text-conditional diffusion model and showed that classifier free guidance has better performance than CLIP [39] guidance. DALLE-2 [40] modified GLIDE to generate semantically consistent images conditioned on a CLIP image embedding, and proposed a diffusion prior that produces the image embedding given a text caption. MDM [46] utilized a classifier-free diffusion-based generative model for text-to-motion and action-to-motion tasks, allowing motion completion and editing as well.

3. Method

Given L segments of audio $A_{1:L} = (A_1, A_2, ..., A_L)$ and a reference image I_{ref} as inputs, our model M synthesizes video frames $\hat{X}_{1:L} = (\hat{X}_1, \hat{X}_2, ..., \hat{X}_L)$ having the same identity as I_{ref} and lip motion synchronized to $A_{1:L}$:

$$\hat{X}_{1:L} = \mathrm{M}(A_{1:L}, I_{ref}). \tag{1}$$

An overview of our proposed framework is shown in Fig. 2, which consists of three major components:

- **Lip & Non-lip Disentanglement** Given a pre-trained identity- and appearance-irrelevant facial motion encoder E_v, we first leverage it to learn two complementary features, $\mathbf{f}_l^a$ for lip, and $\mathbf{f}_{nl}^v$ for non-lip features including pose, expression, blink and gaze.
- **Audio to Non-lip Diffusion Prior** A diffusion prior network P_{a2nl} models the one-to-many mapping from audio feature $\mathbf{f}^a$ to the "hallucinated" non-lip feature $\mathbf{f}_{nl}^a$, which is trained to be close to its visual counterpart $\mathbf{f}_{nl}^v$, allowing audio-only facial driving at inference time.
- **Audio-based Talking Head Generation** Given an identity feature $\mathbf{f}_{id}$, a lip feature $\mathbf{f}_l^a$, and a non-lip feature $\mathbf{f}_{nl}^a$ generated with the diffusion prior, we concatenate them together and feed into a GAN similar to LPD [6] or PC-AVS [65] to output the final reenacted video.

3.1. Lip & Non-lip Disentanglement

The identity encoder E_{id} is designed to capture identity and appearance information, while E_v is instead to remove both while encoding all facial motions. Both encoders are pre-trained in a similar setting as LPD [6] and PC-AVS [65]. After pretraining, we conduct lip & non-lip disentanglement consisting of audio-visual contrastive learning for lip feature learning and decorrelation for non-lip feature learning, as well as face reconstruction, as described below:

Audio-Visual Pretraining for Audio Lip Space An audio encoder E_a is trained to provide accurate audio feature $\mathbf{f}^a = \mathrm{E}_a(A_{1:N})$ through contrastive learning [39] with $\mathbf{f}_{cl}^v = \mathrm{MLP}_{\ cl}(\mathrm{E}_v(X_{1:N}))$, where N is the number of samples and $X_{1:N}$ are N frames in the same video corresponding to audio $A_{1:N}$. Then $\mathbf{f}^a$ is further compressed to filter out lip-irrelevant information and get audio lip feature $\mathbf{f}_l^a$ through MLP_{a2l} since $\mathbf{f}^a$ mainly contains lip related information, as shown in Sec. 4.3. The contrastive loss is defined as $L_{con}(m,n) = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(m_i/\|m_i\|_2 \cdot n_i/\|n_i\|_2)}{\sum_{j=1}^{N}\exp(m_i/\|m_i\|_2 \cdot n_j/\|n_j\|_2)}$, resulting in a total contrastive loss between $\mathbf{f}^a$ and $\mathbf{f}_{cl}^v$,

$$L_{cl} = \frac{1}{2}[L_{con}(\mathbf{f}^a, \mathbf{f}_{cl}^v) + L_{con}(\mathbf{f}_{cl}^v, \mathbf{f}^a)]. \tag{2}$$

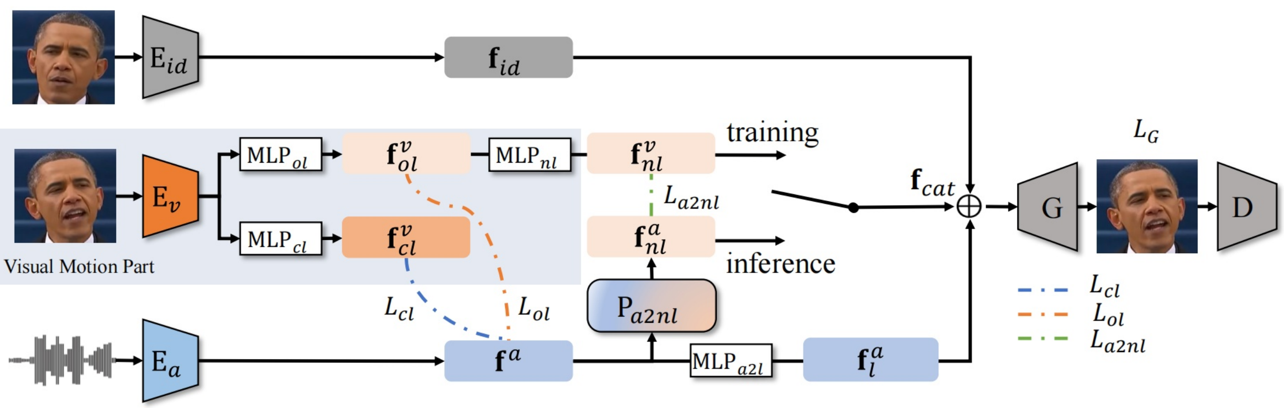

Figure 2. The overall pipeline of our proposed framework. The dotted lines represent the loss functions that are used only in training, *i.e.*, L_{cl}, L_{ol} and L_{a2nl}. L_G is the training loss for visual non-lip space. Note that there is a switch in the figure, which means the different forward processes in training and inference: in training, the concatenated feature $\mathbf{f}_{cat} = cat(\mathbf{f}_{id}, \mathbf{f}^v_{nl}, \mathbf{f}^a_l)$; at inference stage, $\mathbf{f}_{cat} = cat(\mathbf{f}_{id}, \mathbf{f}^a_{nl}, \mathbf{f}^a_l)$. Thus, the visual motion part is not needed anymore at the inference stage.

Reconstruction Learning for Visual Non-lip Space To learn a good non-lip space, we propose an orthogonal loss to penalize the correlation between $\mathbf{f}^v_{ol}$ and $\mathbf{f}^a$, which uses the well-learned audio space to disentangle lip-related motions. The orthogonal loss is used together with reconstruction since there should also be a completeness constraint to avoid mode collapse of lip-irrelevant features. In practice, we maintain two memory banks (**MB**) for storing $\mathbf{f}^v_{ol}$ and $\mathbf{f}^a$ in previous $K-1$ steps, in order to have more samples than N. We denote the feature dimension of $\mathbf{f}^a$ as d_a, and $\mathbf{f}^v_{ol}$ as d_v. The orthogonal loss is defined as follows:

$$L_{ol} = \frac{1}{d_v} \sum_{i=1}^{d_v} \sum_{j=1}^{d_a} P_{cor}(\mathbf{f}^v_{ol}, \mathbf{f}^a)^2_{(i,j)}, \tag{3}$$

where $P_{cor}(\mathbf{f}^v_{ol}, \mathbf{f}^a) \in d_a \times d_v$ computes the Pearson correlation coefficient between $\mathbf{f}^v_{ol}$ and $\mathbf{f}^a$. Since the correlation between lip motions and audio features is strong, L_{ol} (which encourages the linear uncorrelation between $\mathbf{f}^v_{ol}$ and $\mathbf{f}^a$) is effective to disentangle lip-related motions. Then, MLP$_{nl}$ further projects $\mathbf{f}^v_{ol}$ into a smaller space to remove undesired content and get non-lip feature $\mathbf{f}^v_{nl}$.

We follow the reconstruction loss L_{GAN}, L_{L1} and L_{VGG} used in [65]. The generator G receives the concatenation $\mathbf{f}_{cat} = cat(\mathbf{f}_{id}, \mathbf{f}^v_{nl}, \mathbf{f}^a_l)$ as inputs and generates final images I_{gen} with modulated convolution. The discriminator D is trained jointly by generative adversarial learning [23, 12]. Note that we omit the batch-mean operation here for convenience, but in practice, all losses used here are calculated with the average in a batch. Different from [65], we propose an additional gaze loss L_{gaze} by utilizing a pre-trained gaze encoder [1], which calculates the $L1$ distance between the gaze features of generated images and ground-truth (GT) images I as follows:

$$L_{gaze} = \left\| \Phi(I) - \Phi(I_{gen}) \right\|_1 . \tag{4}$$

The total loss L_G is given as:

$$\begin{aligned} L_G = \lambda_{ol} \cdot L_{ol} + \lambda_{gaze} \cdot L_{gaze} + \lambda_{L1} \cdot L_{L1} \\ + \lambda_{\text{GAN}} \cdot L_{\text{GAN}} + \lambda_{\text{VGG}} \cdot L_{\text{VGG}}, \end{aligned} \tag{5}$$

where λ s are weights for losses.

3.2. Audio2nonlip Diffusion Prior

Beside a strong relationship with lip motions, audio features have a complex non-linear relationship with non-lip motions and it is a one-to-many mapping problem. This is due to the fact that there are many reasonable facial motions that can be corresponded to the same audio input. As inspired by DALLE-2 [40] and MDM [46], we design an audio2nonlip diffusion prior network P$_{a2nl}$ to address this problem. Our proposed network is depicted in Fig. 3.

Diffusion Model Diffusion model is designed based on the stochastic diffusion process. The forward diffusion process is defined below:

$$q(n^t_{1:L} | n^{t-1}_{1:L}) = \mathcal{N}(n^{t-1}_{1:L}; \sqrt{\alpha^t} n^{t-1}_{1:L}, (1-\alpha^t)I), \tag{6}$$

In our context, $n_{1:L} = (n_1, n_2, ..., n_L)$ is a sequence of non-lip feature $\mathbf{f}^v_{nl}$. α^t is a hyper-parameter and $t \sim [1, T]$ is the time step of the diffusion process. Eq. 6 can be approximated to $n^T_{1:L} \sim \mathcal{N}(0, 1)$ if α^t is small enough.

During the reversed diffusion process, P$_{a2nl}$ models the audio2nonlip distribution as $p(n^0_{1:L} | a_{1:L})$, where $a_{1:L} = (a_1, a_2, ..., a_L)$ is a sequence of audio feature $\mathbf{f}^a$. Instead of predicting the noise as formulated by vanilla DDPM [45], P$_{a2nl}$ learns to predict the initial signal itself with the simplified objective function [45] described as follows:

$$L_{simple} = \mathbb{E}_{n^0 \sim q(n^0 | \mathbf{f}^a), t \sim [1,T]} [\left\| n^0 - \mathrm{P}_{a2nl}(n^t, t, \mathbf{f}^a) \right\|_2], \tag{7}$$

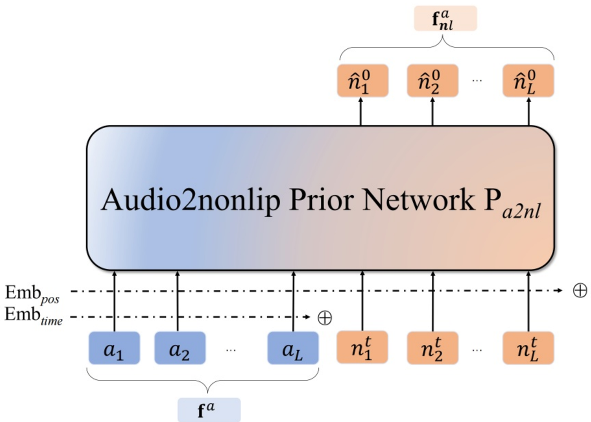

Figure 3. Our **audio2nonlip** diffusion prior network P_{a2nl}.

where we use n^0 to represent $n^0_{1:L}$ for convenience.

We implement P_{a2nl} with a transformer encoder [51] architecture. As shown in Fig. 3, an audio feature $a_{1:L}$ is added to the time embedding Emb_{time}. Then, it's concatenated with noisy non-lip feature $n^t_{1:L}$ and added with positional embeddings Emb_{pos} [51]. The model encodes the embeddings with bidirectional self-attention and feed-forward layers then output denoised non-lip feature $n^0_{1:L}$.

Velocity Loss Note that our designed P_{a2nl} denoises $n^t_{1:L}$ in a non-autoregressive way. To encourage the naturalness and coherence of generated non-lip motion, we borrow the velocity loss from MDM [46] defined as follows:

$$L_{vel} = \frac{1}{L-1}\sum_{i=2}^{L}\left\|(\hat{n}^0_i - \hat{n}^0_{i-1}) - (n^0_i - n^0_{i-1})\right\|_2, \quad (8)$$

where n^0_i is the GT non-lip feature and $\hat{n}^0_i$ is the predicted denoised non-lip feature at the i-th position respectively. The intuition here is that the difference between adjacent non-lip features should be close to the difference in the GT. The total loss of this stage is defined as:

$$L_{a2nl} = L_{simple} + L_{vel}. \quad (9)$$

Classifier-free Guidance We use classifier-free guidance [35] for conditioned diffusion generation. In training time, we randomly set the condition to $\emptyset$ for 10% of the samples so that $\mathrm{P}_{a2nl}(n^t, t, \emptyset)$ approximates n^0. At inference stage, the output of the P_{a2nl} is extrapolated further in the direction of $\mathrm{P}_{a2nl}(n^t, t, \mathbf{f}^a)$ and away from $\mathrm{P}_{a2nl}(n^t, t, \emptyset)$:

$$\mathrm{P}_{a2nl}(n^t, t, \mathbf{f}^a) = s \cdot \mathrm{P}_{a2nl}(n^t, t, \mathbf{f}^a) + (1-s) \cdot \mathrm{P}_{a2nl}(n^t, t, \emptyset), \quad (10)$$

where s is a scaling parameter while increasing it improves sample quality at the cost of diversity.

Sequential Mask Editing To deliver a smooth non-lip facial sequence with arbitrary length, we design a mechanism to ensure continuity between generated segments, and provide an editing method to do so. Specifically, we randomly mask 90% of the non-lip tokens n^0 and concatenate them with noisy non-lip tokens n^t as the input of P_{a2nl} at training stage. Practically, we found it easy for P_{a2nl} to quickly learn how to fill in the masked region using hints from the non-lip features in the unmasked regions, generating continuous feature prediction without any extra design in training loss. As a result, it is possible to use P_{a2nl} for audio-guided non-lip motion generation as well as non-lip motion editing at inference time, providing non-lip facial motion with good naturalism and diversity.

3.3. Audio-based Talking Head Generation

The main difference at inference time is that non-lip visual feature $\mathbf{f}^v_{nl}$ is replaced with $\mathbf{f}^a_{nl}$ generated by P_{a2nl} through reversed diffusion process. Thanks to the random noise introduced in the diffusion process, P_{a2nl} can generate various visual non-lip features given the same audio input, resulting in diverse and reasonable reenactment videos without any extra needs of driving video sources.

To generate a video sequence longer than the maximal length of the model input with smooth transitions, the first input non-lip feature to P_{a2nl} is set to the last generated non-lip feature from P_{a2nl} in the previous step, to ensure continuity in non-lip facial motion.

4. Experiments

Our proposed method was evaluated from two aspects: 1) synchronization between audio and lip motions; 2) naturalness and richness of lip-irrelevant facial motions. Both quantitative and qualitative experiments were conducted to showcase the superiority of our method. In addition, we also come up with a novel metric to measure the overall quality of generated facial motions.

Datasets Our model was trained on VoxCeleb2 and evaluated on both VoxCeleb1 and VoxCeleb2. Here are the details of the two datasets,

- **VoxCeleb1 [33]:** The dataset consists of 100,000 utterances from 1,251 celebrities. 100 videos with 25 identities in total were randomly chosen for the test.
- **VoxCeleb2 [7]:** The dataset contains 1 million utterances with 6,112 identities. 500 test videos with 25 identities were randomly chosen for the test.

Comparing Methods Some prior works such as PC-AVS [65], GC-AVT [28], and EAMM [20] require additional driving sources instead of predicting them based on

Method	A	S	$\mathbf{S}_c^{gt}$	VoxCeleb2			VoxCeleb1		
				FID ↓	S_c ↑	N_c ↓	FID ↓	S_c ↑	N_c ↓
Wav2Lip [38]	✗	✗	7.80	22.3	**9.23**	0.18	44.7	**8.80**	0.13
EAMM [20]	✗	✗	1.76	26.1	4.75	1.70	42.6	2.78	0.58
PC-AVS [65]	✗	✗	7.35	14.4	8.21	0.12	35.4	8.42	0.15
MakeItTalk [66]	✓	✗	7.35	19.5	2.03	0.72	40.2	2.16	0.71
Audio2head [54]	✓	✗	7.35	101	6.42	0.13	104	6.65	0.10
Ours + AR	✓	✗	7.35	24.3	7.23	0.02	49.7	7.05	0.04
Ours + Diff.	✓	✓	7.35	**14.2**	7.34	**0.00**	**35.0**	7.31	**0.01**

Table 1. The quantitative results of synchronization and image quality under self-reenactment scenario. **Bold** means the best. **A** indicates the ability to **generate** non-lip motions with **audio-only** signals without other signals, while **S** means the ability to sample **different** results for each round.

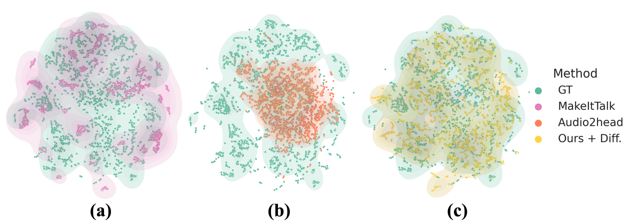

Figure 4. Distribution visualization of generated poses.

audio input. We only compare those methods on lip synchronization and image quality but emphasize that our setup is targeting a more practical scenario. Other works such as Wav2Lip [38] only focus on revising the lip region without touching other facial motions, which are not directly comparable either. MakeItTalk [66] and Audio2head [54] are two closely related works to ours in terms of problem setup, however, our method is designed to predict all the non-lip facial motions with diversity, rather than only limited poses.

Implementation Details Backbones of our models including E_{id}, E_v, E_a, G, and D are borrowed from [65]. For the diffusion prior network P_{a2nl}, we use a 8-layer transformer [51] with 512-d tokens and 1024-d fully feed-forward layers. Note that our models were trained on VoxCeleb2 only but tested on both VoxCeleb1 and VoxCeleb2. All models were trained on 4 NVIDIA A100 GPUs.

4.1. Quantitative Evaluation

Evaluation Metrics We evaluate the performance of generated talking head from the following aspects: image quality, audio-lip sync accuracy, the variation and naturalness of non-lip facial motion. Frechet Inception Distance (**FID**) score [14] is used for the evaluation of image quality. A lower **FID** score indicates a lower distance between the distribution of generated images and real images. Following prior works [38, 65], we use SyncNet Error-Confidence ($\mathbf{S}_c$) as the indicator of audio-lip synchronization, where greater confidence indicates better synchronization. However, the normalized confidence score **NLSE-C** (short for $\mathbf{N}_c$) is used to address the concern raised in recent works [57, 53] regarding the strong relationship between $\mathbf{S}_c$ and its training data, which can make it unfair to compare methods trained on different data. $\mathbf{N}_c$ is defined as $\mathbf{N}_c = \frac{|\mathbf{S}_c^{gen} - \mathbf{S}_c^{gt}|}{\mathbf{S}_c^{gt}}$, where $\mathbf{S}_c^{gen}$ is the confidence value of generated images and $\mathbf{S}_c^{gt}$ is that of its training data.

To evaluate facial motions such as pose, expression, and blink, we utilize a pre-trained 3D morphable face model [9] and include shape irrelevant 3DMM parameters to calculate the following metrics.

- **Var**: The variance of generated facial motions, *i.e.*, the variance of the 3DMM coefficients for each video is calculated and then averaged over the test set. A closer **Var** to GT indicates a better match to the variation of real data.
- $\mathbf{FID}_{fm}$: FID score of 3DMM coefficients calculated as follows: $\mathbf{FID}_{fm} = \frac{1}{K}\sum_{i=1}^{K}\mathbf{FID}(\beta_i)$, where β_i is a sequence of 3DMM coefficients in the i-th video.
- $\mathbf{FID}_{\Delta fm}$: FID score of 3DMM coefficient difference between consecutive frames, *i.e.*, $\mathbf{FID}_{\Delta fm}$, which is similar to $\mathbf{FID}_{fm}$ but the 3DMM coefficient difference between consecutive frames is measured, taking temporal naturalness into consideration.
- **SND**: Our new proposed metric, denoted as **Sequence Naturalness Distance**. It is the sum of $\mathbf{FID}_{fm}$ and $\mathbf{FID}_{\Delta fm}$ indicating the difference of distribution between generated motion and GT motion, from both spatial and temporal perspectives, *i.e.*, $\mathbf{SND} = \mathbf{FID}_{fm} + \mathbf{FID}_{\Delta fm}$. The lower **SND**, the better naturalness.

Evaluation Results Our main results consist of two parts: audio-lip synchronization and image quality as shown in Table 1; richness and sequence naturalness which are shown in Table 2.

It shows that our method with diffusion prior archives the best synchronization ability compared to other methods in Table 1. Note that although Wav2Lip [38] and PC-AVS [65] achieve the highest SyncNet scores, our best N_c indicates the most substantial synchronization ability [57], which is demonstrated in Sec. 4.2. Meanwhile, the image quality of our model with diffusion prior is the best in both test sets.

In Table 2, our method shows significantly higher variance than the other two audio-driven methods and it is closer to the variance of real data. This indicates that our method can produce reasonable diverse head movements and expressions, rather than slight head movement. It's notable that auto-regressive prior, trained with causal attention and regression loss, has a higher variance than GT because it often generates extreme motions, showing a low naturalness score in Table 3. For naturalness, our method achieves the lowest $\mathbf{FID}_{fm}$, $\mathbf{FID}_{\Delta fm}$ as well as **SND** on

Method	VoxCeleb2				VoxCeleb1			
	Var →	FID$_{fm}$ ↓	FID$_{\Delta fm}$ ↓	SND ↓	Var →	FID$_{fm}$ ↓	FID$_{\Delta fm}$ ↓	SND ↓
GT	1.98	-	-	-	1.88	-	-	-
MakeItTalk [66]	0.67	4.70	1.74	6.44	0.80	4.27	1.07	5.34
Audio2head [54]	0.89	5.94	1.30	7.24	0.76	5.03	1.01	6.04
Ours + AR	3.07	5.43	2.22	7.65	2.92	6.21	1.68	7.89
Ours + Diff.	1.57	**3.60**	**1.08**	**4.68**	**1.66**	**3.98**	**0.87**	**4.85**
w/o L_{vel}	**2.09**	4.02	1.27	5.29	2.28	4.65	1.11	5.76
training w/o editing	3.09	6.76	2.00	8.76	3.06	6.96	1.70	8.66

Table 2. The quantitative results of variance and naturalness of **generated** non-lip motions under self-reenactment scenario. "→" indicates a closer score to GT is better. Due to setting differences, *i.e.*, Wav2Lip, EAMM, and PC-AVS, are not valid comparisons here.

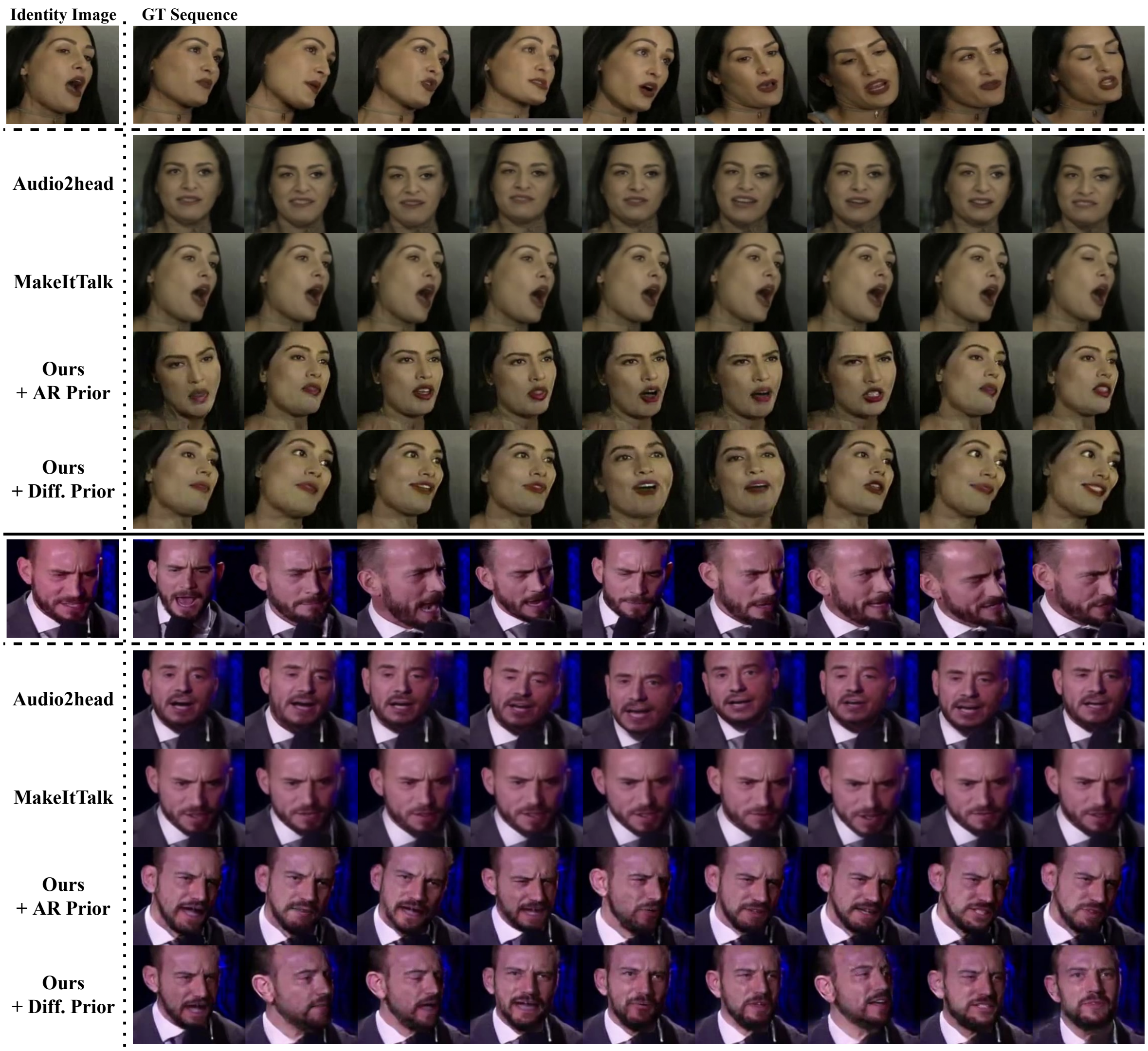

Figure 5. Qualitative results of our method compared to other baselines. Each row shows nine uniformly sampled frames from videos. Here we use two audio sources to drive different identities respectively. Our method shows accurate lip-audio synchronization with diverse and natural poses and expressions.

both VoxCeleb1 and VoxCeleb2, surpassing other state-of-the-art by a large margin. Also, we randomly sample 5,000 poses for each method and employ t-SNE [50] for visualization in Fig. 4, which demonstrates that the distribution we generate is the most comprehensive to GT. More details and examples are shown in the supplementary. In summary,

Method	Sync Accuracy	Naturalness	Richness
GT	4.61	4.41	4.18
MakeItTalk [66]	1.97	2.35	2.10
PC-AVS [65]*	3.13	2.68	2.68
Audio2head [54]	2.56	2.35	2.34
Ours + AR	3.46	2.53	3.37
Ours + Diff.	**4.08**	**3.82**	**3.68**

Table 3. Human evaluation on generated samples on VoxCeleb2. * means that PC-AVS was evaluated on cross-reenactment scenario.

Method	Lip Only		Non-lip Only			
	$S_c \uparrow$	$N_c \downarrow$	$B_d \downarrow$	$G_d \downarrow$	$E_d \downarrow$	$P_d \downarrow$
GT	7.35	-	-	-	-	-
w/o L_{ol}	2.84	0.61	0.0074	0.1960	0.0568	0.0021
w/o **MB**	**7.69**	0.05	0.0112	0.0315	0.0875	0.0293
Ours	7.48	**0.03**	**0.0067**	**0.0163**	**0.0558**	**0.0019**

Table 4. Evaluation of disentanglement between lip and non-lip on VoxCeleb2. B_d and G_d denote L2 distances of 2D landmark [58] of eyes and iris, respectively, while E_d and P_d denote L2 distances of 3DMM coefficients of poses and expressions accordingly. Note that $\mathbf{f}_{nl}^{v}$ is fixed while tested on lip only metrics, and $\mathbf{f}_{l}^{a}$ is fixed while tested on non-lip only metrics for fair comparisons.

our method can generate rich facial motions with excellent realism (*i.e.*, a lower distribution distance) and naturalness.

4.2. Qualitative Evaluation

Two audio-driven video sequences of our method verses other baselines are shown in Fig. 5. It's clearly shown that our method is capable of producing much more natural-looking and diverse facial motions than other baselines, including pose, expression, blink and gaze, while maintaining accurate audio-lip synchronization compared to the GT.

User Study We invited 20 subjects for human evaluations, focusing on three aspects: 1) the accuracy of audio-lip synchronization; 2) the naturalness of non-lip motions and coherency between non-lip motions and audios; 3) the richness of non-lip motions. 50 videos are generated from audio and identity in the test set using the following methods: our diffusion prior, our auto-regressive prior, MakeItTalk [66], Audio2head [54] and PC-AVS [65] (driven by poses from another video). After shuffling the generated videos, each annotator rates from 1 (bad) to 5 (good) according to Mean Opinion Scores (MOS) rating protocol. Table 3 tells that our model with diffusion prior archives the best synchronization, naturalness and richness among all the models. Our model with auto-regressive prior archives relatively good richness compared to the remaining models, we attribute it to the exposure bias problem [4] brought by auto-regressive generation, which leads to high richness but low naturalness. Note that PC-AVS shows incompatible result in Table 3 as compared to Table 1 on $\mathbf{S}_c$, indicating that choosing pose driving signals from a random video may lead to synthesis results with bad audio-lip synchronization.

It's interesting that our proposed **SND** scores reflect the **naturalness** of user study to some degree. Nevertheless, we leave it to future work for rigorous study of their correlation through more human evaluations.

4.3. Ablation

Table 4 discusses the performance of lip & non-lip disentanglement. While tested on lip only metrics without L_{ol}, SyncNet scores get lower, which indicates that G tends to generate mouth movements with information extracted from visual modality instead of audio. Thus, lip & non-lip space are not fully disentangled. Without **MB**, the performance of driving non-lip motions gets worse. Because it is hard to use a large batch to compute correlation, resulting in a not well decoupled non-lip space.

We conducted ablation study for P_{a2nl} from three aspects: 1) auto-regressive prior versus diffusion prior; 2) with/without velocity loss L_{vel}; 3) training with/without editing mechanism. Diffusion prior shows siginificant improvements on naturalness scores including FID_{fm}, $\mathrm{FID}_{\Delta fm}$ and **SND**, as compared to autoregressive prior shown in Table 2, with slightly better synchronization ability in Table 1. Fig. 6 shows that without L_{vel}, the prediction of non-autogressive diffusion prior model P_{a2nl} becomes unstable, which leads to a larger variance and worse **SND** scores in Table 2. While training without the editing mechanism mentioned in Sec. 3.2, we applied editing only during sampling as in MDM [46], and observed that it jitters between adjacent frames as shown in Fig. 7, resulting in a poorer **SND** score in Table 2. Note that our model trained with editing mechanism generates smoother results.

GT

w/o L_{vel}

w/ L_{vel}

Figure 6. Ablation study of our diffusion prior P_{a2nl} w or w/o L_{vel}. Each row shows five uniformly sampled frames.

5. Conclusions

In this paper, we introduce a novel talking head generation method based on a diffusion prior model, which can generate diverse and natural-looking talking head videos with only audio and an identity image as inputs. Such "*audio-driving is all you need*" setting is very friendly

Figure 7. Ablation study of our diffusion prior P_{a2nl} trained w or w/o editing. Each row shows five adjacent frame.

for reenactment and dubbing applications. Comprehensive evaluations including our newly proposed metrics validated the effectiveness of our system.

Limitations Our method mainly focuses on non-lip motion generation. We leave the improvement of rendering quality to future works.

References

[1] Ahmed A Abdelrahman, Thorsten Hempel, Aly Khalifa, and Ayoub Al-Hamadi. L2cs-net: Fine-grained gaze estimation in unconstrained environments. *arXiv preprint arXiv:2203.03339*, 2022. 4

[2] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-Supervised MultiModal Versatile Networks. In *NeurIPS*, 2020. 3

[3] Mohammed M Alghamdi, He Wang, Andrew J Bulpitt, and David C Hogg. Talking head from speech audio using a pre-trained image generator. *arXiv preprint arXiv:2209.04252*, 2022. 2, 3

[4] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28, 2015. 8

[5] Stella Bounareli, Vasileios Argyriou, and Georgios Tzimiropoulos. Finding directions in gan's latent space for neural face reenactment. *arXiv preprint arXiv:2202.00046*, 2022. 3

[6] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 3

[7] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 5

[8] Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang. Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3965–3969. IEEE, 2019. 2

[9] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 6

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3

[11] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. Visual speech enhancement. *arXiv preprint arXiv:1711.08789*, 2017. 2

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 4

[13] Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3

[16] Fangzhou Hong, Liang Pan, Zhongang Cai, and Ziwei Liu. Versatile multi-modal pre-training for human-centric perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16156–16166, 2022. 3

[17] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3397–3406, 2022. 3

[18] Gee-Sern Hsu, Chun-Hung Tsai, and Hung-Yi Wu. Dual-generator face reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 642–650, 2022. 3

[19] Po-Hsiang Huang, Fu-En Yang, and Yu-Chiang Frank Wang. Learning identity-invariant motion representations for cross-id face reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7084–7092, 2020. 3

[20] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. *arXiv preprint arXiv:2205.15278*, 2022. 5, 6

[21] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[22] Venkatesh S Kadandale, Juan F Montesinos, and Gloria Haro. Vocalist: An audio-visual synchronisation model for lips and voices. *arXiv preprint arXiv:2204.02090*, 2022. 2

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 4

[24] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics*, 37(4):163:1–14, Aug. 2018. 3

[25] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020. 3

[26] Jiyoung Lee, Soo-Whan Chung, Sunok Kim, Hong-Goo Kang, and Kwanghoon Sohn. Looking into your speech: Learning cross-modal affinity for audio-visual speech separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1336–1345, 2021. 2

[27] Borong Liang, Yan Pan, Zhizhi Guo, Hang Zhou, Zhibin Hong, Xiaoguang Han, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Expressive talking head generation with granular audio-visual control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3387–3396, June 2022. 2, 3

[28] Borong Liang, Yan Pan, Zhizhi Guo, Hang Zhou, Zhibin Hong, Xiaoguang Han, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Expressive talking head generation with granular audio-visual control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3387–3396, 2022. 5

[29] Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (TOG)*, 40(6):1–17, 2021. 2, 3

[30] Moustafa Meshry, Saksham Suri, Larry S Davis, and Abhinav Shrivastava. Learned spatial representations for few-shot talking-head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13829–13838, 2021. 3

[31] Dongchan Min, Minyoung Song, and Sung Ju Hwang. Styletalker: One-shot style-based audio-driven talking head video generation. *preprint arXiv:2208.10922*, 2022. 3

[32] Gaurav Mittal and Baoyuan Wang. Animating face using disentangled audio representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 2

[33] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017. 5

[34] Arsha Nagrani, Joon son Chung, Samuel Albanie, and Andrew Zisserman. Disentangled speech embeddings using cross-modal self-supervision, 2020. 2

[35] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3, 5

[36] Andrew Zisserman Olivia Wiles, A. Sophia Koepke. X2face: A network for controlling face generation by using images, audio, and pose codes. In *ECCV 2018*, 2018. 3

[37] Se Jin Park, Minsu Kim, Joanna Hong, Jeongsoo Choi, and Yong Man Ro. Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory. In *36th AAAI Conference on Artificial Intelligence (AAAI 22)*. Association for the Advancement of Artificial Intelligence, 2022. 3

[38] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020. 2, 3, 6

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3

[40] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3, 4

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3

[42] Shuai Shen, Wanhua Li, Zheng Zhu, Yueqi Duan, Jie Zhou, and Jiwen Lu. Learning dynamic facial radiance fields for few-shot talking head synthesis. *arXiv preprint arXiv:2207.11770*, 2022. 2

[43] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3

[44] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3

[45] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 4

[46] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 3, 4, 5, 8

[47] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. *ECCV 2020*, 2020. 3

[48] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 3

[49] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 2

[50] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5, 6

[52] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *CoRR*, abs/1906.06337, 2019. 3

[53] Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. *arXiv preprint arXiv:2211.14506*, 2022. 6

[54] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv preprint arXiv:2107.09293*, 2021. 2, 6, 7, 8

[55] Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. One-shot talking face generation from single-speaker audio-visual correlation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2531–2539, 2022. 3

[56] Chao Xu, Jiangning Zhang, Yue Han, Guanzhong Tian, Xianfang Zeng, Ying Tai, Yabiao Wang, Chengjie Wang, and Yong Liu. Designing one unified framework for high-fidelity face reenactment and swapping. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 54–71. Springer, 2022. 3

[57] Shunyu Yao, RuiZhe Zhong, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering. *arXiv preprint arXiv:2201.00791*, 2022. 6

[58] Baosheng Yu and Dacheng Tao. Heatmap regression via randomized rounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 8

[59] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 3

[60] Xianfang Zeng, Yusu Pan, Mengmeng Wang, Jiangning Zhang, and Yong Liu. Realistic face reenactment via self-supervised disentangling of identity and pose. pages 12757–12764. AAAI Press, 2020. 3

[61] Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. Facial: Synthesizing dynamic talking face with implicit attribute learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3867–3876, 2021. 2, 3

[62] Jiangning Zhang, Liang Liu, Zhucun Xue, and Yong Liu. Apb2face: Audio-guided face reenactment with auxiliary pose and blink signals. 2020. 3

[63] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 3

[64] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9299–9306, 2019. 2

[65] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4176–4186, 2021. 2, 3, 4, 5, 6, 8

[66] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020. 2, 3, 6, 7, 8