Diff2Lip: Audio Conditioned Diffusion Models for Lip-Synchronization

Soumik Mukhopadhyay[1] soumik@umd.edu
Saksham Suri[1] sakshams@cs.umd.edu
Ravi Teja Gadde[2] rtg267@nyu.edu
Abhinav Shrivastava[1] abhinav@cs.umd.edu

[1]University of Maryland, College Park [2]New York University

Abstract

The task of lip synchronization (lip-sync) seeks to match the lips of human faces with different audio. It has various applications in the film industry as well as for creating virtual avatars and for video conferencing. This is a challenging problem as one needs to simultaneously introduce detailed, realistic lip movements while preserving the identity, pose, emotions, and image quality. Many of the previous methods trying to solve this problem suffer from image quality degradation due to a lack of complete contextual information. In this paper, we present Diff2Lip, an audio-conditioned diffusion-based model which is able to do lip synchronization in-the-wild while preserving these qualities. We train our model on Voxceleb2, a video dataset containing in-the-wild talking face videos. Extensive studies show that our method outperforms popular methods like Wav2Lip and PC-AVS in Fréchet inception distance (FID) metric and Mean Opinion Scores (MOS) of the users. We show results on both reconstruction (same audio-video inputs) as well as cross (different audio-video inputs) settings on Voxceleb2 and LRW datasets. Video results are available at https://soumik-kanad.github.io/diff2lip.

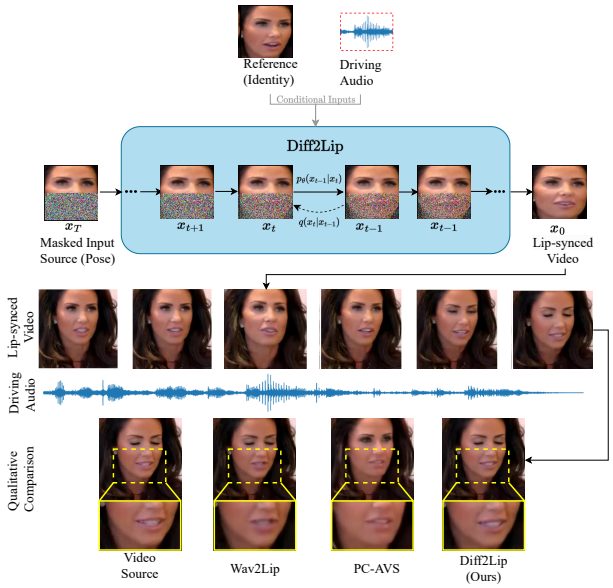

Figure 1. **Top:** Our Diff2Lip approach uses an audio-conditioned diffusion model to generate lip-synchronized videos. (Here q denotes the forward diffusion process and p_θ is the learned reverse diffusion process.) **Bottom:** On zooming in to the mouth region it can be seen that our method generates high-quality video frames without suffering from identity loss.

1. Introduction

Oscar-winning director Bong Joon Ho famously pointed out that subtitles act as a barrier between a foreign audience not adept in the language and their ability to fully enjoy amazing movies [1], as the viewer needs to focus on both watching and reading. A rarely explored alternative, multiple-language version movie (MLV), where the same film is shot in multiple languages in parallel, is naturally much more expensive [49]. While dubbing is a popular compromise solution, it can feel unnatural due to the lack of synchronization between speech and actors' video. As a cheaper alternative, lip-synchronization (lip-sync) aims to generate the mouth region of the human face such that the lips correspond to a different speech audio. Its applications beyond movies include education, virtual avatars, video conferencing, assistive technology, and culture preservation. Ideally, lip-sync should support any identity and audio from unseen sources (in-the-wild). This brings up the challenges of preserving the actors' identity, pose, emotions, and visual quality while maintaining a realistic lip-sync.

One of the earliest lip-sync methods, Video Rewrite [2], had a purpose-built solution by mapping phonemes to mouth shapes and then blending them onto the target video. Modern techniques have more general solutions but suffer certain limitations. For instance, PC-AVS [58] and GC-AVT [28], disentangle pose and expression respectively but fail to preserve identity (Fig. 1 bottom), have worse visual quality, and have border inconsistencies (while putting the generated heads back to the scene). On the other

hand, works that target a specific identity, such as SynthesizingObama [47], require video/identity-specific training. Other methods which rely on extracting intermediate representations, e.g., landmarks in MakeItTalk [59], have to deal with estimation errors in these representations. Finally, approaches that can generalize on in-the-wild lip-sync settings pose it as an inpainting task, where the mouth region is masked and then generated according to the audio. Examples include Wav2Lip [35], which achieves good lip-sync but at the cost of poor visual quality (see Fig. 1 bottom), and AV-CAT [46], which has a multistage pipeline but does not capture finer details. In this paper, we introduce Diff2Lip, an inpainting style approach that solves the lip-sync task using diffusion models, which addresses most of these shortcomings and achieves visually superior lip-sync results.

We propose an audio-conditioned diffusion model to solve the task of lip-sync (Fig. 1 top). Diffusion models [19] are likelihood-based models that can generate astonishing results in high variation datasets (e.g., ImageNet [12]), that GANs [15] cannot match. To generalize in-the-wild, we pose the problem as a conditional diffusion model based inpainting task [39]. Diff2Lip takes three inputs: a masked input frame, a reference frame, and an audio frame, and outputs the lip-synced mouth region. Diff2Lip leverages (1) the masked input frame to get the pose context; (2) the reference frame to get the identity and mouth region textures; (3) the audio frame to drive the lip shape. Using an audio+image conditioned diffusion model, Diff2Lip maintains a fine balance between all these contextual input information, avoiding lip-sync problems (e.g. identity loss, reference copying, inaccurate lip shape). Diff2Lip optimizes three losses: a reconstruction loss to guide synthesis; a sync-expert loss [35] to enforce synchronization; and a sequential adversarial loss to enforce inter-frame continuity. Diff2Lip generates high image quality without identity loss or generalizability issues as shown in Fig. 1 bottom.

We evaluate our work on commonly used benchmarks of Voxceleb2 [8] and LRW [10] datasets for the tasks of reconstruction and cross generation (see section 4). We compare against popular methods used for lip-sync like Wav2Lip [35] and PC-AVS [58]. Extensive evaluations show that Diff2Lip outperform existing methods in terms of image fidelity while having comparable synchronization.

The following are the contributions of this work:

- We propose a novel diffusion model based approach for audio-conditioned image generation.
- Using frame-wise and sequential losses we are able to successfully generate high quality lip-sync.
- We show that the use of a sequential adversarial loss makes frame-wise video generation more stable for diffusion models across frames.
- Extensive evaluations validate that our generations outperform existing methods in FID metric and MOS of the users showing the effectiveness of Diff2Lip.

2. Related Works

In this section, we first talk about existing methods in lip-sync and then discuss conditional diffusion models.

2.1. Lip synchronization

Lip-sync methods can be roughly classified into the following four categories. Please note that there may be overlaps between these categories.

Embedding-based head reconstruction. This class of methods tends to synthesize the entire head by the fusion of speech and identity features. This is usually done using an encoder-decoder style architecture. Song et al. [45] and Vougioukas et al. [50] use RNNs while Speech2Vid [7] uses CNNs. LipGAN [25] uses an audio-visual discriminator to improve synchronization. PC-AVS [58] performs disentanglement of identity, speech, and pose from each other to have complete pose control. GC-AVT [28] additionally disentangles emotion. Recent contemporary works like DiffTalk [41] use latent diffusion models for achieving high visual quality at the cost of lip-sync, which is even worse in cross generation. This method additionally requires landmarks for proper face positioning, uses auto-regressive inference strategies that cannot be parallelized, and employs an external frame interpolation method as it suffers from jitter. In general, as these methods generate full faces, they suffer from border inconsistency issues while putting the generated head back onto the frame.

Intermediate representation-based methods. These methods learn to manipulate sparse intermediate representations like face landmarks or meshes. Chen et al. [6] and Das et al. [11] generate faces conditioned on the landmarks estimated using the audio. MakeItTalk [59] proposes to predict speaker landmark displacement based on the audio. Methods like Song et al. [44], Yu et al. [55], and Xie et al. [54] use 3DMM (facial mesh) to generate face videos. Neural Voice Puppetry [48], uses audio to predict the expression basis coefficients of a 3D model. Zhang et al. [57] propose to first predict 3DMM-based animation parameters which are then converted into a dense flow for facial animation. Although these methods leverage intermediate structures, getting such representations manually is expensive while automatic predictions are error-prone. Further, these tend to lose finer details given the sparse representation.

Personalized methods. In this type of methods, the models are trained to be identity specific or even video-specific [27]. For example, SynthesizingObama [47] only focuses on Obama's lip sync using an audio-to-landmark network. MEAD [24] leverages edges while Lu et al. [30] uses facial landmarks to create edge-like conditional-feature maps to generate talking faces. Methods like Song et al. [44], Zhang et al. [57], and Neural Voice Puppetry [48],

discussed earlier, which do audio-driven expression manipulation (3DMM) also fall in this category. Some methods also deal with explicit 3D mesh vertex deformfation like LipSync3D [26]. NeRF [31] based models like AD-NeRF [16] and SSPNeRF [29] are also person-specific. This class of methods demonstrates high video quality at times, but that comes at the cost of retraining the model on the specific person and environment every time.

Inpainting-based methods. In these methods instead of generating the whole face only the bottom part of the face, which gets affected by speech is modified. These models don't suffer from image boundary inconsistencies when pasting back the mouth portion to the entire frame. Initial works like [5] only focus on lip region features and used an audio-speech fusion module to merge them. Wav2Lip [35] is one of the most popular methods in this area and shows the importance of lip-sync expert network for better lip-sync. AV-CAT [46] uses a transformer backbone and a refinement model for inpainting the lower face. SyncTalkFace [34] further introduces an audio lip memory that is used for inference time generation. Our method falls also in this category of lip-sync. It doesn't struggle with error propagation issues being end-to-end trainable, and does not require any explicit 2D/3D information while being identity-agnostic. It further improves the image quality compared to previous methods. Recent contemporary works like Gupta et al. [17] also focus on improving quality by using VQ-GAN and a face restoration network but consequently make the lips have a pinkish tint and slightly different from the input. Their lip-sync expert network requires five times more context compared to Wav2Lip.

2.2. Conditional diffusion models

Initial work in this area involved class conditioning for image generation. For example, Ho & Salimans [20] used class labels to train a diffusion model by interpolating between conditional and unconditional outputs. While Guided-diffusion [13] used a classification network for class conditioning to get a better image generation. Methods like GLIDE [32], DALLE-2 [36], Stable-Diffusion [37], and IMAGEN [40] leverage language models to generate photorealistic as well as many other styles of images just using text prompt inputs. There has also been some research into text-to-video generation like Video Diffusion Models [21] and more photorealistic models like Gen-1 [14]. Recently, seeing the popularity of diffusion models, people have also proposed models like Noise2Music [22] that can generate music using just text prompts. There has also been some work on image-conditioned diffusion models. Palette [39] is a generalized image-to-image generation framework, which can solve tasks like coloration, inpainting, outpainting, jpeg-restoration, etc. We also pose the problem as an inpainting style diffusion task but with additional audio and reference identity-conditioned inputs.

3. Methods

In this section, we discuss our proposed approach - Diff2Lip. We propose a novel audio and image conditioned diffusion model which is able to synthesize high quality lip-synced mouths corresponding to the audio input. We discuss diffusion models in Section 3.1. Then we introduce our approach in Section 3.2. Finally, in Section 3.2.1, we talk about the losses required to train our model.

3.1. Diffusion Models

Diffusion models [19] are likelihood-based models which try to sample points from a given distribution by gradually denoizing random gaussian noise in T steps. In the forward diffusion process, increasing amounts of noise is added to a sample point x_0 iteratively as $x_0 \rightarrow x_1 \rightarrow \cdots \rightarrow x_{t-1} \rightarrow x_t \rightarrow \cdots \rightarrow x_{T-1} \rightarrow x_T$, to get a completely noisy image x_T. Formally, the forward diffusion process is a Markovian noising process defined by a list of noise scales $\{\bar{\alpha}_t\}_{t=1}^{T}$ as:

$$q(x_t|x_0) := \mathcal{N}(x_t|\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{1}$$

which can be rewritten as:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \epsilon \in \mathcal{N}(0, \mathbf{I}) \tag{2}$$

where ϵ is the noise, $\mathcal{N}$ denotes normal distribution, x_0 is the original image, and x_t is noised image after t steps of the diffusion process. The reverse diffusion process aims to learn the posterior distribution $q(x_{t-1}|x_0, x_t)$, using which one can estimate x_{t-1} given x_t. This is typically done using a neural network, which can be parameterized in multiple ways. Similar to [13, 19, 33], we choose to parameterize the neural network to predict the noise, ie. $\epsilon_\theta(x_t, t)$, where θ represents the parameters of the neural network. It takes a noisy sample x_t and timestep t to predict the added noise ϵ in Eq. 2. The model is learned using the simplified objective used in [19] which reweights the variational lower bound on the maximum likelihood objective:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{x_0,t,\epsilon}[\|\epsilon_\theta(x_t, t) - \epsilon\|_2^2] \tag{3}$$

The posterior distribution $q(x_{t-1}|x_0, x_t)$ is also tractable using the Bayesian rule and turns out to be another normal distribution. When using DDIM [43] for sampling, we can deterministically sample the posterior by disregarding the variance. Since we can write x_0 in terms of x_t and ϵ using Eq. 2, therefore we can recover x_{t-1} deterministically given x_t and ϵ using:

$$x_{t-1} = \sqrt{\frac{\bar{\alpha}_{t-1}}{\bar{\alpha}_t}}x_t + \left(\sqrt{1-\bar{\alpha}_{t-1}} - \sqrt{\frac{\bar{\alpha}_{t-1}(1-\bar{\alpha}_t)}{\bar{\alpha}_t}}\right)\cdot\epsilon \tag{4}$$

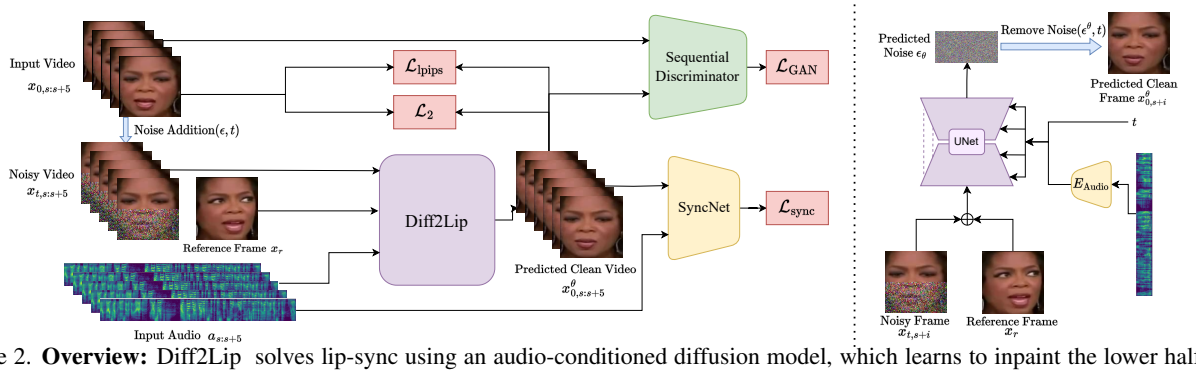

Figure 2. **Overview:** Diff2Lip solves lip-sync using an audio-conditioned diffusion model, which learns to inpaint the lower half of the face. During training (left), given an input video sequence $x_{0,s:s+5}$, we first add noise to the lower half using the forward process (Eq. 2) to get the noisy video sequence $x_{t,s:s+5}$, where diffusion step t is sampled uniformly. Then a noisy video frame $x_{t,s+i}$ for $i \in [0, 5)$, a different random reference frame x_r, and the audio frame a_{s+i} is input to our model. The audio encoder E_{Audio} encodes the audio frame a_{s+i}. Our model (right), predicts the added noise ϵ_θ given these inputs, which is used to get the predicted clean frame $x^{\theta}_{0,s+i}$ (using Eq. 2). Then frame-wise reconstruction losses like $\mathcal{L}_2$ and $\mathcal{L}_{\text{lpips}}$ are applied to the predicted clean sequence $x^{\theta}_{0,s:s+5}$ for enforcing good image quality while sequential losses like sequential adversarial loss $\mathcal{L}_{\text{GAN}}$ and SyncNet expert loss $\mathcal{L}_{\text{sync}}$ ensure lip-sync.

This equation represents the mean of the learned posterior $p_\theta(x_{t-1}|x_{x_t})$ distribution in the DDIM [43] formulation.

For sampling during inference time, x_T is sampled from the standard normal distribution. The neural network can then recover the noise ϵ_θ that needs to be removed. This in turn can be fed into Eq. 4 to get back x_{T-1}. Iterating over this one can get the clean image as $x_T \rightarrow x_{T-1} \rightarrow \cdots \rightarrow x_t \rightarrow x_{t-1} \rightarrow \cdots \rightarrow x_1 \rightarrow x_0$ as seen in Fig. 3 top.

Notation. In this paper, we work with diffusion processes and videos. We use t for the diffusion process step number while s for the video frame number. For the diffusion process, we keep the notation here the same as [33].

3.2. Proposed Approach

We pose the problem of lip-sync as a lower mouth inpainting task, where given an input face with the lower half masked, an audio frame input, and a reference frame input, the model needs to generate the masked region of the face. Formally, given a video $V = \{v_1, \ldots v_S\}$ with S frames, were v_s is the s^{th} frame, and audio $A = \{a_1, \ldots, a_S\}$, where a_s is the s^{th} audio frame, our model processes one video frame $x_{0,s} = v_s$ at a time. Let $x_{s,T} = v_s \cdot (1 - M) + \eta \cdot M$ be a noise-masked video frame, where $\eta \in \mathcal{N}(0, \mathbf{I})$ and M is a binary mask for the lower half of the face. (Here the subscript T denotes a completely noised frame that we want to denoise). We want our trained model to be able to recover v_s using the reverse diffusion process, given inputs masked video frame $x_{s,T}$, the audio frame a_s, and a random reference frame $x_r = v_{\text{random}(1,S) \neq s}$. This setup is quite similar to Wav2Lip [35]. The random reference frame x_r is chosen from the same video and provides cues about the source's identity and pose. We make sure that it is not the same as the input frame; otherwise there could be information leakage while training. The audio input a_s provides information about the lip structure.

Figure 3. Intermediate x_t (**top**) and x^{θ}_0 (**bottom**) as t goes from T to 0 (left to right), sampled at uniform intervals.

As shown Fig. 2 we formulate the problem as an inpainting task similar to [39], i.e., we learn a conditional model $\epsilon_\theta(x_{s,t}, a_s, x_r, t)$. At training time, we take a clean sample frame $x_{s,0} (= v_s)$ and a uniformly sampled t, and add noise to $x_{s,0}$ using Eq. 2 to get $x_{s,t}$. The model is trained to predict the noise $\epsilon \in \mathcal{N}(0, \mathbf{I})$ added to it using Eq. 3.

We feed the reference frame by concatenating it with the input frame while the audio is fed using group normalization (similar to time and class conditioning in [33]). Our network has a UNet [38] backbone which consists of residual blocks and attention blocks similar to [13]. We want the UNet to extract contextual information from the unmasked portion of the input frame, and the reference frame. To enforce this we provide these directly as input to the network. For the audio which is used as a conditioning, we first encode it using a trainable encoder E_{Audio}, which generates embeddings that are injected as conditioning. E_{Audio} is also built using the same blocks as the UNet.

3.2.1 Additional Losses

When just training using $\mathcal{L}_{\text{simple}}$ (applied to the masked region), we observe that the mouth region generation had good image quality but no lip-sync. Hence we add additional losses to make our model work.

Our model predicts in noise space and hence many image-space losses cannot be directly applied to it. There are three ways to approach this issue - first, our model could be parameterized to directly predict the denoized image x_0

instead of predicting ϵ. Second, we can use the sampling process described in Section 3.1 to recover back the clean image x_0. Third, substituting x_t and ϵ_θ into Eq. 2 one could directly recover $x_0^\theta(x_t, t)$, an estimate of x_0, without having to do iterative sampling. We observed that directly predicting denoized image leads to worse image quality while using iterative sampling is overly time-consuming and hence we stick with predicting ϵ.

This approach leads to a noisy $x_0^\theta(x_t, t)$ when the step t is large as seen in Fig. 3 bottom but there have been previous works [51] which have applied image losses to $x_0^\theta(x_t, t)$. We enforce an $\mathcal{L}_2$ loss on this $x_0^\theta(x_t, t)$ to make it clean:

$$\mathcal{L}_2 = \mathbb{E}_{x_{0,s},t,\epsilon}[\|x_{0,s}^\theta - x_{0,s}\|_2^2] \tag{5}$$

Next, to impose audio synchronization, we use SyncNet discriminator as used by Wav2Lip [35]. We first separately train the SyncNet in a contrastive manner which is kept fixed during training our generation model. Similar to Wav2Lip [35] we work with a sequence of 5 frames as input to the SyncNet. By using 5 predicted video frames $x_{0,s:s+5}^\theta$ and the corresponding audio sequence $a_{s:s+5}$, SyncNet loss can be written as:

$$\mathcal{L}_{\text{sync}} = \mathbb{E}_{x_{0,s},t,\epsilon}[\text{SyncNet}(x_{0,s:s+5}^\theta, a_{s:s+5})] \tag{6}$$

As shown in our ablation, directly adding SyncNet loss, deteriorates the image quality. To mitigate this we add perceptual similarity loss [56] on the generated frames:

$$\mathcal{L}_{\text{lpips}} = \mathbb{E}_{x_{0,s},t,\epsilon}\mathbb{E}_l[\|\phi_l(x_{0,s}^\theta) - \phi_l(x_{0,s})\|_2^2] \tag{7}$$

where $\phi_l(\cdot)$ represents the features coming from the l^{th} layer of a pretrained-VGG network. Finally, to enforce temporal consistency we also add a sequence adversarial loss. This makes the movement of the lips realistic across frames.

$$\begin{aligned}\mathcal{L}_{\text{GAN}} = \mathbb{E}_{x_{0,s},t,\epsilon}[\log D_\psi(x_{0,s:s+5}^\theta)] + \\ \mathbb{E}_{x_{0,s}}[\log(1 - D_\psi(x_{0,s:s+5}))]\end{aligned} \tag{8}$$

where we use a PatchGAN [23] discriminator D_ψ. This task requires more context than just two frames [4] but no optical flow [52]. The overall optimization objective can be written as:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{simple}} + \lambda_{\text{l2}}\mathcal{L}_2 + \lambda_{\text{sync}}\mathcal{L}_{\text{sync}} + \\ \lambda_{\text{lpips}}\mathcal{L}_{\text{lpips}} + \lambda_{\text{gan}}\mathcal{L}_{\text{GAN}}\end{aligned} \tag{9}$$

For sequence-based losses, it is essential that the diffusion process step input t is the same for a sequence of frames $x_{0,s:s+5}$. This ensures uniformity within a predicted sequence during loss computation.

Table 1. Ablation over our losses (Reconstruction)

Losses	FID ↓	SSIM ↑	PSNR ↑	LMD ↓	Sync_c ↑
Reconstruction	8.589	0.523	18.234	3.472	0.633
+ SyncNet	8.998	0.526	**18.57**	3.123	6.336
+ Perceptual	**7.751**	0.526	18.548	3.121	6.53
+ Seq. GAN	8.213	**0.527**	18.52	**3.101**	**7.89**

4. Experiments

Datasets. We evaluate our method on the Voxceleb2 [8] and LRW [10] datasets, which contain in-the-wild videos of talking human faces and are commonly used for lip-sync research.

Voxceleb2 [8] - consists of over 1M face-cropped Youtube videos coming from 6000+ identities. This dataset consists of high variation in lighting, image quality, pose, and motion blur. The average video length is 8 seconds.

LRW [10] - is a lip-reading dataset that contains 1000 videos each of 500 different words for a length of 1 second coming from BBC news. It has less variation compared to Voxceleb2 and focuses on clean front-facing videos. Like previous works, our model is trained only on the Voxceleb2 train split while we test on both datasets. We don't use the whole dataset for training but rather only use the first utterance of every video, which totals 145K videos.

Implementation Details. We preprocess the videos to have a framerate of 25 fps and an audio sample rate of 16kHz. For all our models the video resolution is 224×224 out of which we crop the face and resize it to 128×128. This is then masked in the lower half using gaussian noise and fed to our model which only morphs the lower half of this image according to the audio input. Then we resize it back to the original crop size and place it back on the video. For audio inputs, we first sample the audio at 16kHz and then create mel-spectrograms with window-size 800 and hop-size 200. These audio frames turn out to have size 16×80. We build our code on top of the guided-diffusion repository [13]. We train our model on 8 NVIDIA RTXA6000 GPUs which takes around 4 days. Our model is trained using $T = 1000$ diffusion steps, but for faster inference, we use only 25 steps of DDIM [43] sampling which takes 4.67 seconds on an average for all the frames of one VoxCeleb2 [8] video (avg. 8 seconds at 25 fps) on 8 NVIDIA RTXA6000 GPUs.

Comparison Methods. We compare our method against the most popular methods for lip-sync. Our choice of models is based also on models/codebases which are publicly available. Wav2Lip [35] is an inpainting style method that uses SyncNet expert loss to get good lip-sync. PC-AVS [58] is a head reconstruction method that focuses on controlling pose apart from identity and lip shape. For both these methods we use their publicly available pre-trained models for the evaluation of all the datasets.

Table 2. Quantitative comparison with baselines on Voxceleb2 [8] and LRW [10] on the task of reconstruction and Cross generation.

Dataset	Method	Reconstruction						Cross			
		FID ↓	SSIM ↑	PSNR ↑	LMD ↓	$Sync_c$ ↑	$Sync_d$ ↓	FID ↓	LMD ↓	$Sync_c$ ↑	$Sync_d$ ↓
VoxCeleb2	Wav2Lip [35]	3.26	0.53	18.18	3.16	**9.08**	5.93	5.11	4.84	**8.12**	6.74
	PC-AVS [58]	4.25	0.53	**18.26**	3.16	6.71	7.80	10.62	5.00	6.96	7.53
	Diff2Lip (Ours)	**2.46**	0.53	18.09	**3.04**	8.78	**5.93**	**4.53**	**4.82**	7.62	**6.73**
LRW	Wav2Lip [35]	4.23	**0.68**	**20.76**	**2.15**	**8.13**	**6.09**	5.19	**3.88**	**7.52**	**6.56**
	PC-AVS [58]	6.80	0.61	20.10	2.29	6.68	7.29	8.48	4.09	6.66	7.27
	Diff2Lip (Ours)	**2.62**	0.67	20.62	2.17	7.41	6.21	**2.54**	3.93	6.44	6.97

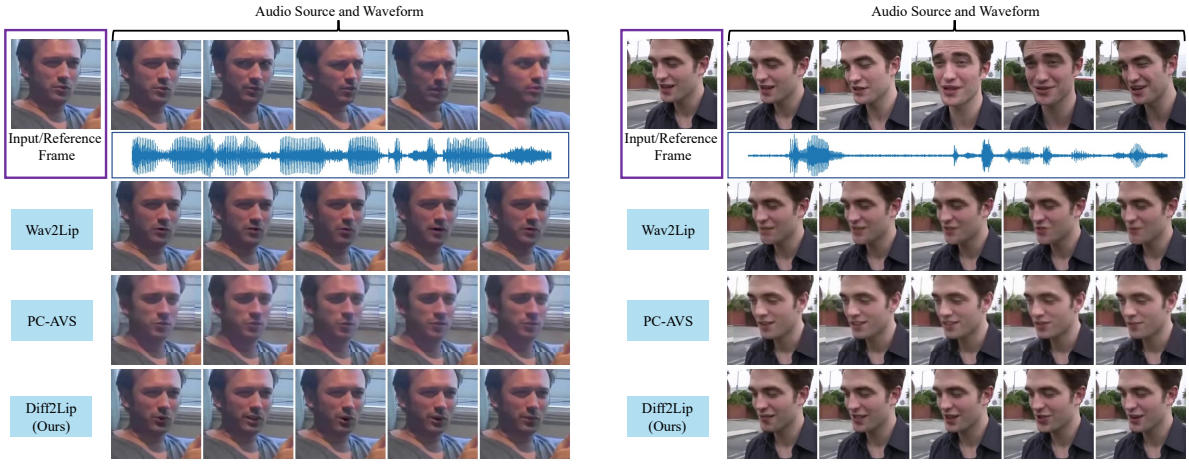

Figure 4. **Qualitative results of Reconstruction on VoxCeleb2 [8]**. Here we provide only the first frame as the input source (for pose) as well as the reference frame (for identity), and this frame is driven using the audio (second row) coming from the same video (top row). Wav2Lip [35] blurs the lip region in both cases to achieve the correct lip shape while PC-AVS has identity loss (see right) and border discontinuity. Our generations look highly realistic and have the lip shapes as in the audio source. (Please zoom in for better visibility.)

Table 3. Ablation over our losses (Cross generation)

Losses	FID ↓	LMD ↓	$Sync_c$ ↑
Reconstruction	6.694	5.313	0.992
+ SyncNet	8.784	**4.816**	5.946
+ Perceptual	5.016	5.009	5.955
+ Seq. GAN	**4.592**	4.985	**6.83**

4.1. Quantitative Evaluation

For quantitative evaluation, we evaluate our model in terms of both the visual quality as well as audio synchronization. For visual quality, we use FID [18], SSIM [53], and PSNR, which are popular metrics used in papers like Wav2Lip [35], PC-AVS [58], AV-CAT [46]. FID is a popular metric used for comparing the "realness" of generated images by comparing against the real image distribution. SSIM and PSNR are pixel-wise image similarity metrics that compare a pair of images and are not suited for capturing variability in video generation [42] but are included in this work for completeness. While to measure synchronization we use LMD [5], $Sync_c$, and $Sync_d$ [9]. LMD measures the distance between mouth landmarks among frames. $Sync_c$ is the confidence score of SyncNet while $Sync_d$ is the average distance between SyncNet video and audio representations, which tell the synchronization quality. Note that for evaluation, we use the pre-trained SyncNet from the SyncNet's [9] repository but for training, we train our own SyncNet, similar to Wav2Lip [35] and AV-CAT [46]. We use pre-trained models of Wav2Lip and PC-AVS methods to conduct our evaluations. Wav2Lip has provided its code for calculating FID and $Sync_c$ and we use the same for these metrics. We use face-alignment [3] for landmark detection and the LMD metric proposed in [5]. For SSIM and PSNR we use the same inputs as used for FID, consequently, our values are a bit different compared to [58]. This is possibly because they evaluate these metrics at different scales and lack these evaluation details. On the other hand, we tend to get better LMD and $Sync_c$ values for PC-AVS than noted in their paper. Some papers like [28] show PC-AVS's metrics only for reference as its generations occasionally fail in landmark detection. For uniformity of scale, we uncropped PC-AVS's generations and paste them back on the background before evaluation.

4.1.1 Reconstruction

Similar to the setting mentioned in [58] and later used in [28] and [46], we evaluate the methods on the task of re-

Figure 5. **Qualitative results of Cross generation on VoxCeleb2 [8]**. Here we provide a video source (first row) and drive that identity-pose combination using audio coming from a different video (second and third rows). Wav2Lip [35] blurs its generations, for example, beard region details are missing on the right. PC-AVS's [58] generations have flaws like identity loss, in both cases. They introduce artifacts near the eyes on the left while there is identity loss on the right. Our method generates realistic mouths with expressive lips while being in sync with the audio source. In the bottom 2 rows, we can see that the lip region of our generations match those of the audio source. (Please zoom in for better visibility.)

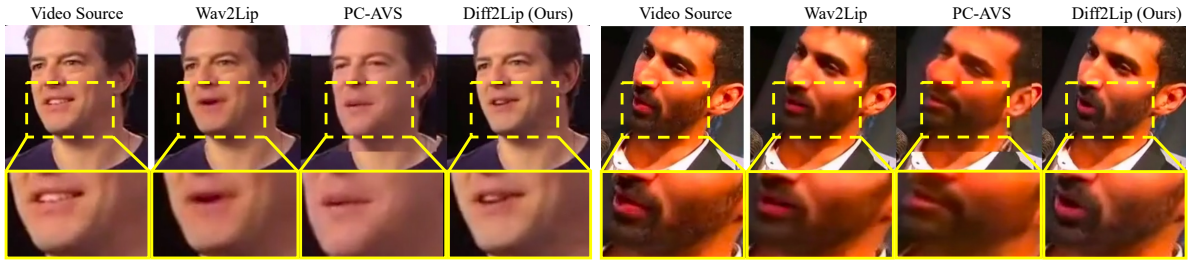

Figure 6. **Qualitative Visual-quality Comparison**. We zoom in on the mouth region of two examples and compare them against the video source. Wav2Lip [35] blurs the lip region while PC-AVS [58] tends to change the identity. Diff2Lip preserves identity and generates high-fidelity lips.

construction of the video given only the first frame and the audio corresponding to the same video. The audio is used as the driver for this reconstruction. Note that PCAVS requires an additional pose input apart from the identity input. We feed the first frame to it for both inputs, similar to the "fixed pose" setting in their paper. For Wav2lip and ours, the first frame acts as both the input frame and the reference frame. For Voxceleb2, we show the results on 4911 test utterances instead of 35K videos due to resource constraints. These 4911 utterances are the first utterance of each video in the test set, and hence cover all the videos in the test set. For LRW, our results are noted on all 25K videos in the test set.

The results are noted in Table 2. We observe that Diff2Lip outperforms both the methods with respect to FID metric on both datasets which points towards better generation quality. The SSIM, PSNR and LMD values of our method are comparable with the other methods. We see that Wav2Lip's Sync_c is better than both ours and PC-AVS's. This is possibly because our SyncNet expert may be weaker in performance compared to the one used in Wav2Lip.

4.1.2 Cross generation

We also evaluate the method on the task of lip-sync when the identity and the pose are controlled using a video while the lip-sync is driven using input audio corresponding to a different video. This was introduced in [35] and is a more realistic setting as here the generations are closer to lip-sync in-the-wild. We use the input frame as the reference frame in this setting similar to Wav2Lip, as that provides the best texture information of the frame. For PC-AVS, the input frames are fed as both pose and identity sources. Note that

Table 4. User Study measured by Mean Opinion Scores (MOS) (max. 5) and Preference in percentage.

Measure	Wav2Lip	PC-AVS	Diff2Lip
MOS (Visual quality) ↑	3.75	2.71	**4.16**
MOS (Lip-sync quality) ↑	3.84	3.34	**3.86**
MOS (Overall quality) ↑	3.70	2.91	**3.91**
Preference ↑	37%	8.33%	**54.67%**

we cannot evaluate SSIM and PSNR for this setting because there are no ground truth frames available. So, we provide the rest of the metrics in Table 2. For Voxceleb2, we select 4970 pairs of audio-video combinations where the two sources are different. We sample these using all the pair combinations of the first utterance of the first video coming from 71 randomly chosen test identities. For LRW, we use the 28K audio-video pair provided in Wav2Lip's [35] evaluation. The results are noted in Table 2. Similar to reconstruction evaluation, we here as well observe that our method excels in image quality while being comparable in other metrics except for Sync_c.

4.2. Qualitative Evaluation

For qualitative evaluation, we show visually compare against on both reconstructions (in Fig. 4) as well as cross generation (in Fig. 5). These settings are the same as introduced in Section 4.1.1 and 4.1.2. It can be observed in these qualitative results that PC-AVS tends to lose the identity of the source video and also suffers from boundary discontinuity problems which make it unsuitable for in-the-wild generation. On the other hand, Wav2Lip tends to generate blurred-out mouth regions so as to achieve good lip sync. Diff2Lip does not suffer from these issues and is able to generate high-quality mouth region while having expressive lip shapes which correctly correspond to the ground truth (audio sources' mouth shape) as seen in Fig. 6.

User Study. We conduct a user study where we ask 15 participants to judge lip-sync videos generated in cross generation setting by Diff2Lip and two other methods. 20 videos were sampled from the VoxCeleb2's test set and are driven by randomly selected driving audios. The participants rated the videos 1-5 (where higher is better) in the aspects of (1) **Visual quality** (2) **Lip-sync quality**, and (3) **Overall quality**. We used the Mean Opinion Score (MOS) measure to aggregate these ratings. Further, we record the percentage of times users preferred a method. We present the results in Table 4, where we see that our method surpasses others in all the categories. In terms of Lip-sync quality, this is opposed to our quantitative results, especially Sync_c. We speculate that Sync_c, might favor blurry generations with high temporal consistency while humans prefer high fidelity over slight temporal inconsistency.

4.3. Ablations

We conduct an ablation study to showcase the contribution of various losses used during training. Specifically, we train our model in three different settings in which we introduce an additional loss in each setting. First, we train our model using only $\mathcal{L}_{\text{simple}}$ (Reconstruction). Second, we train another version of the model using $\mathcal{L}_{\text{simple}} + \mathcal{L}_2 + \mathcal{L}_{\text{sync}}$ (+ SyncNet). Here intuitively the $\mathcal{L}_{\text{sync}}$ should introduce better synchronization. Third, we further add a perceptual loss $\mathcal{L}_{\text{sync}}$ (+ Pecept). We add this loss because adding the SyncNet loss led to worse image quality. Finally, we add the sequential adversarial loss $\mathcal{L}_{\text{GAN}}$ to achieve even temporal consistency(+ Seq. GAN). We test these on a smaller subset of 500 VoxCeleb2 test audio-video pairs in the cross generation setting as well as the reconstruction setting.

It can be seen in Table 3 that moving from "Reconstruction" to "+ SyncNet" gives a sudden improvement in the Sync_c metric. This supports our intuition that only reconstruction-based losses are not enough. We also see that this transition deteriorates the image quality. This gets solved as we move to the "+ Perceptual" setting. Finally, the addition of sequential adversarial loss not just further improves the image quality but also improves the Sync_c, clearly showing the advantage of this loss. In the reconstruction setting in Table 1, most of these observations still hold except FID being lower for "+ Perceptual" than "+ Seq. GAN". This could be attributed to the static nature of the input source in this setting while the ground truth is moving.

5. Discussion and Conclusion

Discussion. Even though Diff2Lip, cannot be used for facial reenactment and hence cannot be used for harmful acts like face-swapping, it can be used for other malicious purposed like disinformation. We discourage and disapprove of any such applications which may have negative implications and strictly condone their use for positive purposes.

Conclusion. In this work, we present Diff2Lip, which is able to generate high-quality lip synchronization. We pose the task to be a mouth region inpainting task and solve it by learning an audio-conditioned diffusion model. Our ablation studies show that SyncNet loss is required in our framework to introduce lip-sync while sequential adversarial loss improves both image quality and temporal consistency. Finally, extensive quantitative and qualitative results validate that our method performs better than state-of-the-art methods in terms of image quality while maintaining other metrics and also being preferred by the users.

Acknowledgements. We would like to thank our colleagues Matthew Gwilliam, Archana Swaminathan, and Ahmed Taha for their feedback, and all the user study participants for their time. This project was partially funded by the DARPA SemaFor (HR001119S0085) and DARPA SAIL-ON (W911NF2020009) programs.

References

[1] Commentary: 'The 1-inch-tall barrier of subtitles': Bong Joon Ho rightly calls out Hollywood myopia, Jan. 2020. 1

[2] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 353–360, 1997. 1

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 6

[4] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5933–5942, 2019. 5

[5] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European conference on computer vision (ECCV)*, pages 520–535, 2018. 3, 6

[6] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generationwith dynamic pixel-wise loss. *arXiv preprint arXiv:1905.03820*, 2019. 2

[7] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? *arXiv preprint arXiv:1705.02966*, 2017. 2

[8] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 2, 5, 6, 7

[9] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016. 6

[10] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In Shang-Hong Lai, Vincent Lepetit, Ko Nishino, and Yoichi Sato, editors, *Computer Vision – ACCV 2016*, pages 87–103, Cham, 2017. Springer International Publishing. 2, 5, 6

[11] Dipanjan Das, Sandika Biswas, Sanjana Sinha, and Brojeshwar Bhowmick. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 408–424. Springer, 2020. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3, 4, 5

[14] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 3

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

[16] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021. 3

[17] Anchit Gupta, Rudrabha Mukhopadhyay, Sindhu Balachandra, Faizan Farooq Khan, Vinay P Namboodiri, and CV Jawahar. Towards generating ultra-high resolution talking-face videos with lip synchronization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5209–5218, 2023. 3

[18] 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. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 6

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3

[21] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 3

[22] Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*, 2023. 3

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 5

[24] 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/CVF conference on computer vision and pattern recognition*, pages 14080–14089, 2021. 2

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

[26] Avisek Lahiri, Vivek Kwatra, Christian Frueh, John Lewis, and Chris Bregler. Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2755–2764, 2021. 3

[27] Lincheng Li, Suzhen Wang, Zhimeng Zhang, Yu Ding, Yixing Zheng, Xin Yu, and Changjie Fan. Write-a-speaker:

Text-based emotional and rhythmic talking-head generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1911–1920, 2021. 2

[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. 1, 2, 6

[29] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-aware implicit neural audio-driven video portrait generation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 106–125. Springer, 2022. 3

[30] 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

[31] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[32] 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

[33] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171. PMLR, 18–24 Jul 2021. 3, 4

[34] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2062–2070, 2022. 3

[35] 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, 4, 5, 6, 7, 8

[36] 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

[37] 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

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4

[39] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2, 3, 4

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 3

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

[42] Gaurav Shrivastava and Abhinav Shrivastava. Diverse video generation using a gaussian process trigger. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 6

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 4, 5

[44] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want. *IEEE Transactions on Information Forensics and Security*, 17:585–598, 2022. 2

[45] Yang Song, Jingwen Zhu, Dawei Li, Xiaolong Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. *arXiv preprint arXiv:1804.04786*, 2018. 2

[46] Yasheng Sun, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Zhibin Hong, Jingtuo Liu, Errui Ding, Jingdong Wang, Ziwei Liu, and Koike Hideki. Masked lip-sync prediction by audio-visual contextual exploitation in transformers. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2, 3, 6

[47] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 2

[48] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 716–731. Springer, 2020. 2

[49] Ginette Vincendeau. Hollywood Babel: The Multiple Language Version. *Screen*, 29(2):24–39, 03 1988. 1

[50] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven facial animation with temporal gans. *arXiv preprint arXiv:1805.09313*, 2018. 2

[51] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. In *arXiv*, 2022. 5

[52] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018. 5

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

[54] Tianyi Xie, Liucheng Liao, Cheng Bi, Benlai Tang, Xiang Yin, Jianfei Yang, Mingjie Wang, Jiali Yao, Yang Zhang, and Zejun Ma. Towards realistic visual dubbing with heterogeneous sources. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1739–1747, 2021. 2

[55] Lingyun Yu, Jun Yu, Mengyan Li, and Qiang Ling. Multimodal inputs driven talking face generation with spatial–temporal dependency. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(1):203–216, 2020. 2

[56] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[57] 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. 2

[58] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 5, 6, 7

[59] 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