DiT-Head: High Resolution Talking Head Synthesis using Diffusion Transformers

anonymous

Keywords: Talking Head Synthesis, Diffusion Transformers.

Abstract: We propose a novel talking head synthesis pipeline called "DiT-Head," which is based on diffusion transformers and uses audio as a condition to drive the denoising process of a diffusion model. Our method is scalable and can generalise to multiple identities while producing high-quality results. We train and evaluate our proposed approach and compare against existing methods of talking head synthesis. We show that our model can compete with these methods in terms of visual quality and lip-sync accuracy. Our results highlight the potential of our proposed approach to be used for a wide range of applications including virtual assistants, entertainment, and education. For a video demonstration of results and our user study, please refer to our supplementary material.

1 INTRODUCTION

Talking head synthesis is a challenging task that aims to generate realistic and expressive faces that match the speech and identity of a given person. In recent years, there has been a growing interest in the development of talking head synthesis models due to their potential applications in media production, virtual avatars and online education. However, current state-of-the-art models struggle with generalising to unseen speakers and limited visual quality. Most existing methods focus on person-specific talking head synthesis (Chen et al., 2020; Doukas et al., 2021; Guo et al., 2021; Shen et al., 2022; Zhang et al., 2022; Ye et al., 2023) and rely on expensive 3D structural representations or implicit neural rendering techniques to improve the performance under large pose changes, but these methods still have limitations in preserving the identity and expression fidelity. Moreover, most methods require a large amount of training data for each identity, which limits their generalisation ability.

With the rapid speed of innovation in machine learning, we have better-performing models for image-based tasks such as Latent Diffusion Models (LDMs) (Rombach et al., 2021) that can generate novel images of high fidelity from text with minimal computational cost and Vision Transformers (ViTs) (Dosovitskiy et al., 2020) that can capture global and local features of images and learn effectively from large-scale data. We argue that we should leverage these models to address the challenges of talking head synthesis in order to achieve higher fidelity and generalisation. We propose a novel talking head synthesis model based on Diffusion Transformers (DiTs) (Peebles and Xie, 2022) that takes a novel audio as a driving condition. Our model exploits the powerful cross-attention mechanism of transformers (Vaswani et al., 2017) to map audio to lip movement and is designed to improve the generalisation performance and the visual quality of synthesised videos. Additionally, DiTs are highly scalable, which can make them cost-efficient and time-efficient for a variety of tasks and dataset sizes.

The proposed model has the potential to significantly enhance the performance of synthesised talking heads in a wide range of applications, including virtual assistants, entertainment, and education. In this paper, we present the main methods and technical details of our model, as well as the experimental results and evaluation.

To summarise, we make the following contributions:

- We design an LDM (Rombach et al., 2021) that substitutes the conventional UNet architecture (Ronneberger et al., 2015) with a ViT (Dosovitskiy et al., 2020; Peebles and Xie, 2022) that can scale up, handle multiple types of conditions and leverage the powerful cross attention mechanism of incorporating conditions. Audio is used to drive the denoising process and thus makes the talking head generation audio driven.
- Facial contour information is preserved through using a polygon-masked (Ngo et al., 2020) ground-truth frame and a reference frame as additional conditions. This lets the network focus on learning audio-facial relations and not back-

ground information.

- The proposed DiT-Head model can generalise to multiple identities with high visual quality and quantitatively outperforms other methods for talking head synthesis in visual quality and lip-sync accuracy.

The remainder of the paper is organised as follows: **Section 2** reviews related work on talking head synthesis and explains the fundamentals of LDMs and DiTs; **Section 3** details our methodology; **Section 4** reports our results; **Section 5** discusses our findings, limitations, future work, and ethical issues, and **Section 6** concludes the paper.

2 Related Work

Talking Head Synthesis Models Talking head synthesis is the process of generating a realistic video of a person's face speaking, based on a driving input. This technology has many potential applications in areas such as virtual assistants, video conferencing, and entertainment. In recent years, machine learning models have been developed to improve the quality of audio-driven talking head synthesis. These models can be broadly divided into two categories: 2D-based and 3D-based methods.

2D-based methods (Zakharov et al., 2019; Zhou et al., 2019; Chen et al., 2019; Zhou et al., 2020; Prajwal et al., 2020) use a sequence of 2D images of a person's face to synthesise a video of the person speaking a given audio. Wav2Lip (Prajwal et al., 2020) is a GAN-based talking head synthesis model that can generate lip-synced videos from audio. This method uses a lip-sync discriminator (Chung and Zisserman, 2016) to ensure that the generated videos are accurately synchronised with the audio inputs. Wav2Lip uses an audio feature, a reference frame from the input video (a distant frame from the same video and identity), and a masked ground truth frame as inputs. This approach excels at producing accurate lip movements for any person, but the visual quality of the generated videos is suboptimal. This is because there is a trade-off between lip-sync accuracy and visual quality. Furthermore, GAN-based models suffer from mode collapse and unstable training (Wang et al., 2022). More similar and concurrent to our work, DiffTalk (Shen et al., 2023) uses a UNet-based LDM to produce videos of a talking head. Smooth audio and landmark features are used to condition the denoising process of the model to produce temporally coherent talking head videos.

3D-based methods (Chen et al., 2020; Doukas et al., 2021; Guo et al., 2021; Shen et al., 2022; Zhang et al., 2022; Ye et al., 2023), on the other hand, use a 3D model (Blanz and Vetter, 1999) of a person's face to synthesise a video of the person speaking the target audio. Meta Talk (Zhang et al., 2022) is a 3D-based method which uses a short target video to produce a high-definition, lip-synchronised talking head video driven by arbitrary audio. The target person's face images are first broken down into 3D face model parameters including expression, geometry, illumination, etc. Then, an audio-to-expression transformation network is used to generate expression parameters. The expression of the target 3D model is then replaced and combined with additional face parameters to render a synthetic face. Finally, a neural rendering network (Ronneberger et al., 2015; Isola et al., 2017)translates the synthetic face into a talking face without loss of definition. Other 3D-based methods use Neural Radiance Fields (NERFs) (Mildenhall et al., 2021) for talking head synthesis (Guo et al., 2021; Shen et al., 2022). NERFs for talking head synthesis work by modelling the 3D geometry and appearance of a person's face and rendering it under different poses and expressions. They can produce more natural and realistic talking videos as they capture the fine details and lighting effects of the face and can handle large head rotations and novel viewpoints as they do not rely on 2D landmarks or warping (Guo et al., 2021; Shen et al., 2022).

Both 2D-based and 3D-based methods have their advantages and disadvantages. 2D-based methods are computationally less expensive and can be trained on smaller datasets, but may not produce as realistic results as 3D-based methods (Wang et al., 2022). 3D-based methods, on the other hand, can produce more realistic results, but are computationally more expensive and require larger datasets for training (Wang et al., 2022). Furthermore, 3D-based methods rely heavily on identity-specific training, and thus do not generalise across different identities without further fine-tuning (Guo et al., 2021; Shen et al., 2022; Wang et al., 2022).

Latent Diffusion Models (LDMs) LDMs (Rombach et al., 2021) are a class of deep generative models that learn to generate high-dimensional data, such as images or videos (Yu et al., 2023; Blattmann et al., 2023), by iteratively diffusing noise through a series of transformation steps and then training a model to learn the reverse process. Given real high-dimensional data x, an encoder E (Esser et al., 2021) can be used to learn a compressed representation of x to generate latent representation z. Using maximum likelihood estimation, LDMs can learn complex latent distributions and generate high-quality samples from

them (Rombach et al., 2021; Lovelace et al., 2022; Liu et al., 2023).

LDMs are versatile and can be applied to various tasks such as image and video synthesis, denoising, and inpainting. These type of generative model also have the advantage of having a single loss function and no discriminator, which makes them more stable and avoid GANs' issues such as mode collapse and vanishing gradient (Wang et al., 2022). Moreover, LDMs outperform GANs in sample quality and mode coverage (Rombach et al., 2021; Dhariwal and Nichol, 2021). However, LDMs have a slow inference process because they need to iteratively run the reverse diffusion process on each sample to eliminate the noise.

Diffusion Transformers (DiTs) Diffusion Transformers (DiTs) are a recent type of generative model that blend the principles of LDMs and transformers. DiTs replace the commonly used UNet architecture in diffusion models with a ViT and can perform better than prior diffusion models on the class conditional ImageNet 512×512 and 256×256 benchmarks (Peebles and Xie, 2022). Given an input image, ViTs divide the image into patches and treat them as tokens for the transformer that are then linearly embedded. This allows the ViT to learn visual representations explicitly through cross-patch information interactions using the self-attention mechanism (Vaswani et al., 2017).

Using transformers, DiTs model the posterior distribution over the latent space, which allows the model to capture intricate correlations between the latent variables. In contrast, standard LDMs use the UNet architecture (Ronneberger et al., 2015) to model the same correlations. Transformers have some advantages over UNets in certain tasks that require modelling long-range contextual interactions and spatial dependencies (Cordonnier et al., 2019). They can leverage global interactions between encoder features and filter out non-semantic features in the decoder by using self- and cross-attention mechanisms (Vaswani et al., 2017; Cordonnier et al., 2019). On the other hand, UNets are based on convolutional layers which are characterised by a limited receptive field and an equivariance property with respect to translations (Thome and Wolf, 2023). In contrast, self-attention layers of ViTs allow handling of long-range dependencies with learned distance functions. Furthermore, ViTs are highly scalable which assists in handling a variety of image tasks of different complexities and diverse datasets. These powerful mechanisms make the ViT an intriguing architecture for use in multi-modal tasks such as talking head synthesis.

Conditioning Mechanisms of Diffusion Transformers Deep learning models often use conditioning mechanisms to incorporate additional information (Rombach et al., 2021; Mirza and Osindero, 2014; Cho et al., 2021). These mechanisms allow models to learn from multiple sources of information and can improve performance in many tasks. DiTs can incorporate conditional information in multiple ways and can model conditional distributions as $q(z|c)$, where c represents the conditional information.

One such conditioning mechanism is concatenation, which involves appending the additional information to the input data. Although effective in natural language processing tasks like sentiment analysis, machine translation, and language modelling, concatenation is less useful when dealing with high-dimensional data of different modalities or unaligned conditional information. A more effective conditioning mechanism is that of cross-attention (Vaswani et al., 2017). The cross-attention mechanism of transformers is a way of computing the relevance between two different sequences of embeddings. For example, in a vision-and-language task, the two sequences can be image patches and text tokens. Cross-attention calculates a weighted sum of the input data, where the weights depend on the similarity between the input data and the additional information. This mechanism has shown to be effective in multi-modal conditioning tasks e.g. text-to-image and text-to-video synthesis (Rombach et al., 2021; Blattmann et al., 2023).

The aim of this work was to produce a viable approach to high-quality, person-agnostic, audio-driven talking head video generation. We made use of the powerful self- and cross-attention mechanisms of DiTs in order to achieve this.

3 Methodology

3.1 Data Pre-processing and Overview

We employed a 2-stage training approach and an additional post-processing step at inference-time to produce temporally coherent lip movements using a DiT and incorporating audio features as a condition. For data pre-processing, we used a face-detection method (Zhang et al., 2017) to locate and crop the face from each frame of the input video. This face crop is then resized to $H \times W \times 3$ where $H \in \mathbb{N}$ and $W \in \mathbb{N}$. Next, facial landmark information was extracted from these images (Bulat and Tzimiropoulos, 2017) where face images $x \in \mathbb{R}^{H \times W \times 3}$. Finally, a convex hull mask (Ngo et al., 2020) was created over a copy of these images using facial landmark information for boundary

definition where masked images $x_m \in \mathbb{R}^{H\times W\times 3}$. This design choice greatly impacted the quality of generated videos as it further localised the learning power of the self- and cross-attention mechanism used to learn the relation between the jaw, lips and audio and not the irrelevant information such as the neck, collar etc. Additionally, in order to assist in lip and lower-mouth generation and blending back into the ground-truth video, we performed Gaussian α-blending to the polygon mask. i.e. the boundaries of the polygon mask were smoothed using a Gaussian kernel. For input audio pre-processing, we sampled at 16kHz and normalised in the range $[-1,1]$.

In the first stage, two autoencoders were trained to faithfully reconstruct the ground-truth images and images where the mouth is masked. In the second stage, a DiT is trained whereby masked and reference frames are used in addition to an audio feature as conditions to drive the denoising process. The addition of masked and reference frames make the process more controllable and generalisable across different identities without requiring additional fine-tuning (Prajwal et al., 2020; Shen et al., 2023). Thanks to the learning mode of the latent space, the model can achieve high-resolution image synthesis with minimal computational costs.

In the next sections, we will provide details of the proposed DiT-Head pipeline including all stages. Figure 1 shows a visual overview of our architecture.

3.2 Stage 1: Latent Feature Generation

Taking our masked and ground-truth images, we trained two VQGANs (Esser et al., 2021) to retrieve the latent codes of the images with masks (E_m) and ground-truth images (E). Firstly, vector quantisation (Van Den Oord et al., 2017) was applied to the continuous latent space which involves mapping continuous latent vectors to discrete indices, which are then used to represent the compressed latent space. The decoder for the unmasked images, (D) is only used for the final output and its weights and those of the encoders are fixed during DiT training. We reduced the dimensionality of the input face image by encoding it into a latent representation with a smaller H and W by a factor f where $H/h = W/w = f$. This made the model learn faster and with less resources. The input face image $x \in \mathbb{R}^{H\times W\times 3}$ can be encoded into $z_{gt} = E(x) \in \mathbb{R}^{h\times w\times 3}$ if it is unmasked, or $z_m = E_m(x) \in \mathbb{R}^{h\times w\times 3}$ if it is masked.

3.3 Stage 2: Conditional Diffusion Transformer

In LDMs, a forward noising process is assumed which transforms sample z_0 of the latent space to a noise vector through a series of T steps of diffusion. The diffusion process is defined by a Markov chain (Geyer, 1992) that starts from the data distribution $q_0(z)$ and ends at a simple prior distribution $q_T(z)$, such as a standard Gaussian. At each step t, the latent variable z_0 is corrupted by adding Gaussian noise $\varepsilon_t \sim \mathcal{N}(0,\beta_t I)$, where β_t is a noise level that increases with T steps (Ho et al., 2020). The full forward process can be formulated as,

$$q(z_{1:T}|z_0) := \prod_{z=1}^{T} q(z_t|z_{t-1}), \tag{1}$$

where the Gaussian noise transition distribution at each step t is given by:

$$q_t(z_t|z_{t-1}) := \mathcal{N}(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t I) \tag{2}$$

where $\beta_t \in (0,1)$ and $1-\beta_t$ represent the hyperparameters of the noise scheduler. As defined by (Ho et al., 2020), using the rule of Bayes and the Markov assumption, we can write the latent variable z_t as:

$$q(z_t|z_0) := \mathcal{N}(z_t; \sqrt{\bar{\alpha}_t}z_0, (1-\bar{\alpha}_t)I), \tag{3}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, and $\alpha_t = 1-\beta_t$. Then, the reverse process $q(z_{t-1}|z_t)$ can be formulated as another Gaussian transition (Xu et al., 2023):

$$q_\theta(z_{t-1}|z_t) := \mathcal{N}(z_{t-1}; \mu_\theta(z_t,t), \sigma_\theta(z_t,t))), \tag{4}$$

where μ_θ and σ_θ denote parameters of a neural network ε_θ. In our work, the DiT was used as ε_θ to learn the denoising objective such that we optimise:

$$L_{DiT} = \mathbb{E}_{z,\varepsilon\sim N(0,1),t}\left[\|\varepsilon - \varepsilon_\theta(z_t,t)\|_2^2\right] \tag{5}$$

where the added noise ε is predicted given a noised latent z_t where z_t is the result of applying forward diffusion to z_0 and $t \in [1,...,T]$. $\tilde{z_t} = z_t - \varepsilon_\theta(z_t,t)$ is the denoised version of z_t at step t. The final denoised version $\tilde{z_0}$ is then mapped to the pixel space with the pre-trained image decoder $\tilde{x} = D(\tilde{z_0})$ where $\tilde{x} \in \mathbb{R}^{H\times W\times 3}$ is the restored face image.

Since we want to train a model that can create a realistic talking head video with the mouth in sync with the audio and the original identity is matching, audio features were used to condition the DiT.

Audio Encoding To incorporate audio information as a condition for talking head synthesis, we made use of a pre-trained Wav2Vec2 model (Baevski et al., 2020). Wav2Vec2 is a deep neural network model

for speech recognition and processing, designed to learn representations directly from raw audio waveforms without requiring prior transcription or phonetic knowledge. The architecture is based on a combination of convolutional neural networks and transformers, and is trained using a self-supervised learning approach (Baevski et al., 2020).

In our approach, for a given input video, the preprocessed audio was converted to audio features using Wav2Vec2 (Baevski et al., 2020). From this audio and using the pre-trained model, individual features can be extracted for each frame of the video. We concatenated 11 audio features per frame to create an audio feature window of 11 (each frame at time-step t is associated with features from 5 frames prior and 5 frames after). Each concatenated feature window represents 0.44s of speech if the input video is sampled at 25 FPS. By concatenating audio information this way, we provided audio-visual temporal context. After introducing audio signals as a condition, the objective can be re-formulated as:

$$L_{DiT} = \mathbb{E}_{z,\varepsilon\sim N(0,1),a,t}\left[\|\varepsilon - \varepsilon_\theta(z_t, a, t)\|_2^2\right] \quad (6)$$

where the concatenated audio features for the i-th frame are denoted as a 363-dimensional vector, $a_i \in \mathbb{R}^{363}$. This audio feature is then further encoded using a linear transformation to $a_h \in \mathbb{R}^{hidden}$ where *hidden* is the hidden dimension.

Person-agnostic Modelling Our goal was to create a model that can produce realistic lip movements for various people and speakers. To achieve this, we used a reference image as an input to our model, following the approach of Wav2Lip (Prajwal et al., 2020). The reference image x_r contains information about the person's face, head orientation and background. It is a random face image of the same person as the target frame, but from a different segment of the video (at least 60 frames away). We also used a masked image x_m as another input to guide the model for mouth region inpainting without relying on the real lip movements.

However, since DiT works in the latent space, we used E and E_m to retrieve the latent representations of x_r and x_m respectively where $z_r = E(x_r) \in \mathbb{R}^{h\times w\times 3}$ and $z_m = E_m(x_m) \in \mathbb{R}^{h\times w\times 3}$. Our final optimisation objective can be re-formulated as:

$$L_{DiT} = \mathbb{E}_{z,\varepsilon\sim N(0,1),z_t,z_r,z_m,a,t}\left[\|\varepsilon - \varepsilon_\theta(z_t, z_r, z_m, a, t)\|_2^2\right] \quad (7)$$

Conditioning Implementation In order to integrate the conditioning information to the DiT, we used concatenation and cross-attention. We concatenated the spatially aligned z_r and z_m and noised input latent z_T channel-wise. This was used as the input for the first layer of the DiT and introduced the masked ground-truth and reference frame as conditions to drive the inpainting process resulting in $z_T \in \mathbb{R}^{h\times w\times 9}$. Furthermore, the audio condition, a and time-step embedding, t were concatenated channel-wise and then introduced to the DiT via cross-attention. In our model, the spatially aligned concatenated latents, z_T serve as the query, while the concatenated audio feature a and time-step embedding t served as the key and value during cross-attention (Vaswani et al., 2017).

3.4 Evolving Inference and Output

A Denoising Diffusion Implicit Model (DDIM) (Song et al., 2020) was used as the sampling method to perform denoising during inference. This method of sampling is faster than the original Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020) method. While both DDPM and DDIM are based on the diffusion process, they differ in their approach to modelling it. DDPM uses a neural network to model the data distribution and reduce noise step-by-step while DDIM learns the mapping from the data to the solution of a time-dependent partial differential equation that models the diffusion implicitly. Overall, DDIM can produce high-quality samples using less denoising steps (Song et al., 2020).

When generating a talking head video during inference, we provided the exact same conditioning as in training for the first frame (a masked image, a random reference image and a concatenated audio feature) however the ground-truth image latent z is substituted with noise. In addition, for subsequent frames, we use the generated noise latent at time-step t of the previous frame, z_{i-1}, as the reference latent z_r as proposed by (Shen et al., 2023; Bigioi et al., 2023). In this way, the generated face of the previous frame is used to provide temporal contextual information for the generation of the next frame and results in a smoother evolution between frames at the output. During inference, we use 250 DDIM steps and the denoised output from the DiT for the i-th frame, $\tilde{z_i} \in \mathbb{R}^{h\times w\times 3}$, is decoded through the unmasked image decoder D to obtain output face image $\tilde{x_i} = D(\tilde{z_i}) \in \mathbb{R}^{H\times W\times 3}$.

3.5 Stage 3: Post-processing

We applied video frame-interpolation (VFI) on the resulting video in order to alleviate temporal jitter around the mouth region. This temporal jitter is com-

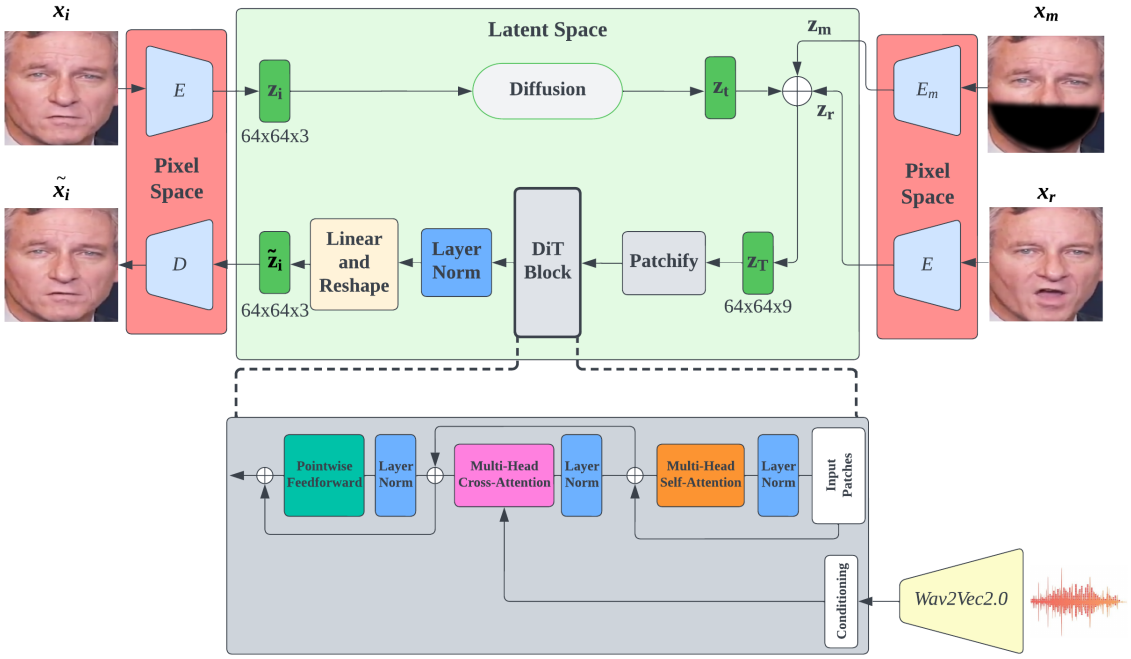

Figure 1: **The DiT-Head architecture**. We train a diffusion-based model for generalised talking head video synthesis and use a DiT to learn the denoising process. The latent representations of the ground-truth image x_i, reference image x_r and masked ground-truth x_m are extracted using VQGAN encoders E and E_m. The latent representations of the reference image z_r and masked ground-truth z_m are concatenated with noise at time-step t, z_t, to produce intermediate representation z_T. Additionally, we utilise the powerful cross-attention mechanism of transformers to introduce audio as a condition.

mon when attempting to iteratively generate temporally coherent images using diffusion-based models (Blattmann et al., 2023). VFI is a technique that can generate intermediate frames between two consecutive frames in a video, resulting in a smoother and more realistic motion. We performed 2x interpolation using the RIFE (Real-time Intermediate Flow Estimation) (Huang et al., 2020) model which uses a neural network that can directly estimate the intermediate flows from images, without relying on bi-directional optical flows that can cause artifacts on motion boundaries (Huang et al., 2020).

3.6 Training Details

6 hours of randomly selected footage from the HDTF dataset (Zhang et al., 2021), containing high-quality videos of diverse speakers, facial expressions and poses, was re-sampled to 25 FPS and used for training. We randomly shift the mask and landmarks by a few pixels during training to improve the model's generalisation ability and use a Gaussian kernel of size 27×27 and σ of 5 for our mask. Our VQGANs (Esser et al., 2021) use a downsampling factor of $f = 4$, produce an intermediate embedding of $z \in \mathbb{R}^{\frac{H}{f} \times \frac{W}{f} \times 3}$ and use a learnable codebook of $N_c \times 3$. In our experiments, $H = 256$, $W = 256$, $N_c = 16384$ and our pre-trained Wav2Vec2 (Baevski et al., 2020) model was *wav2vec2-large-xlsr-53-english*. For the diffusion process, we use the same hyperparameters from ADM (Dhariwal and Nichol, 2021). Our learning rate was constant at 1×10^{-4} and our model had a patch size of 2, 24 DiT blocks, 16 attention heads and a hidden dimension of 1024 (DiT-L/2 (Peebles and Xie, 2022)). The two VQGANs and DiT were trained for 72 hours each using 4 A100 GPUs.

4 Results

4.1 Evaluation Details

Our model was evaluated quantitatively on 11 unseen identities (around 18 minutes of video at 25 FPS) from our randomly selected HDTF (Zhang et al., 2021) test set using ground-truth driving audio. Peak Signal-to-Noise Ratio (PSNR) (Kotevski and Mitrevski, 2010), Structural Similarity Index Measure (SSIM) (Wang et al., 2004), Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018), Fréchet Inception Distance (FID) (Heusel et al., 2017), LSE-C (Lip Sync Error - Confidence) (Prajwal et al., 2020) and LSE-D (Lip Sync Error -

Distance) (Prajwal et al., 2020) are used as quantitative metrics. PSNR measures the pixel error between the original and reconstructed image (Kotevski and Mitrevski, 2010). Higher PSNR means better quality. SSIM measures the similarity between the original and reconstructed image by comparing their structure, brightness and contrast (Wang et al., 2004). Higher SSIM means more similarity. LPIPS measures the perceptual similarity between two images by comparing their features from a deep network (Zhang et al., 2018). It aims to quantitatively reflect how humans perceive images (Zhang et al., 2018). FID measures the realism of generated images by comparing their distributions with ground-truth images using their features from an Inception network (Heusel et al., 2017). Lower scores for LPIPS and FID mean better quality. LSE-C and LSE-D are based on SyncNet (Chung and Zisserman, 2016), which is a lip-sync scorer.

We compared padded face-crop videos from DiT-Head to padded face-crop videos from other talking head synthesis methods such as MakeitTalk (Zhou et al., 2020) and Wav2Lip (Prajwal et al., 2020). This reduces background bias but adds a padded area for blending assessment. MakeitTalk (Zhou et al., 2020) and Wav2Lip (Prajwal et al., 2020) were chosen as they are among the current state-of-the-art for 2D-based person-agnostic talking head synthesis with public implementations available. These implementations were used to generate the output videos. Table 1 presents our quantitative findngs, which includes a quantitative comparison between MakeItTalk (Zhou et al., 2020), Wav2Lip (Prajwal et al., 2020) and DiT-Head w/o Stage 3. We use DiT-Head w/o Stage 3 for a fair comparison between frames and because the lip-sync scorer (Chung and Zisserman, 2016) used for LSE-C(Prajwal et al., 2020) and LSE-D (Prajwal et al., 2020) metrics operates only on 25 FPS videos. Our qualitative findings are presented in Figure 2, which provides a visual comparison between MakeItTalk (Zhou et al., 2020), Wav2Lip (Prajwal et al., 2020) and DiT-Head.

4.2 Quantitative Comparisons with Other Methods

Among the three methods, Table 1 shows that Wav2Lip (Prajwal et al., 2020) performs better than DiT-Head in lip-sync, as it has a lower LSE-D (Prajwal et al., 2020), which measures the lip shape error on the whole region. DiT-Head w/o Stage 3 has a higher PSNR (Kotevski and Mitrevski, 2010) and SSIM (Wang et al., 2004) in addition to lower LPIPS (Zhang et al., 2018) and FID (Heusel et al., 2017) than Wav2Lip. However, DiT-Head also has a lower LSE-C (Prajwal et al., 2020), which measures the lip shape error on the center region. MakeItTalk (Zhou et al., 2020) has the lowest PSNR (Kotevski and Mitrevski, 2010), SSIM (Wang et al., 2004) and LSE-C (Prajwal et al., 2020) as well as the highest LSE-D (Prajwal et al., 2020), LPIPS (Zhang et al., 2018) and FID (Heusel et al., 2017) among the three methods.

4.3 Qualitative Comparisons with Other Methods

Figure 2 shows that MakeItTalk (Zhou et al., 2020) fails to preserve the pose of the input video and results in incorrect head movements and facial alignment, which are especially noticeable in the second identity. MakeItTalk (Zhou et al., 2020) also produces outputs with little expression, as it cannot capture the subtle changes in the facial muscles that convey emotions. On the other hand, Wav2Lip (Prajwal et al., 2020) achieves the most accurate lip shape and expression among the three methods. However, Wav2Lip (Prajwal et al., 2020) suffers from low quality and blurriness around the mouth region. Moreover, Wav2Lip (Prajwal et al., 2020) generates a bounding box around the lower-half of the face, which can be seen as a sharp edge in some cases. This is more noticeable in Wav2Lip (Prajwal et al., 2020) on the second identity compared to DiT-Head. DiT-Head can generate high-resolution outputs with smooth transitions between the generated and original regions. However, DiT-Head has less accurate lip shape and expression than Wav2Lip (Prajwal et al., 2020). Furthermore, it can be seen that both Wav2Lip (Prajwal et al., 2020) and DiT-Head produce less fine-detailed texture of the face when compared to the ground-truth however the lip colour is more accurate in Wav2Lip (Prajwal et al., 2020). We recognise that lip-sync quality is highly subjective. Hence, we encourage readers to view videos in our supplementary material (http://bit.ly/48MgiEr) for a more accurate reflection of the qualitative differences and the findings of our user study on visual quality, lip-sync quality and overall quality of DiT-Head compared to MakeItTalk (Zhou et al., 2020) and Wav2Lip (Prajwal et al., 2020).

5 Discussion

We aimed to compare different methods for talking head synthesis and to propose a novel method that can generate high-resolution and realistic outputs. The quantitative and qualitative results show that our model is competitive against other methods in terms

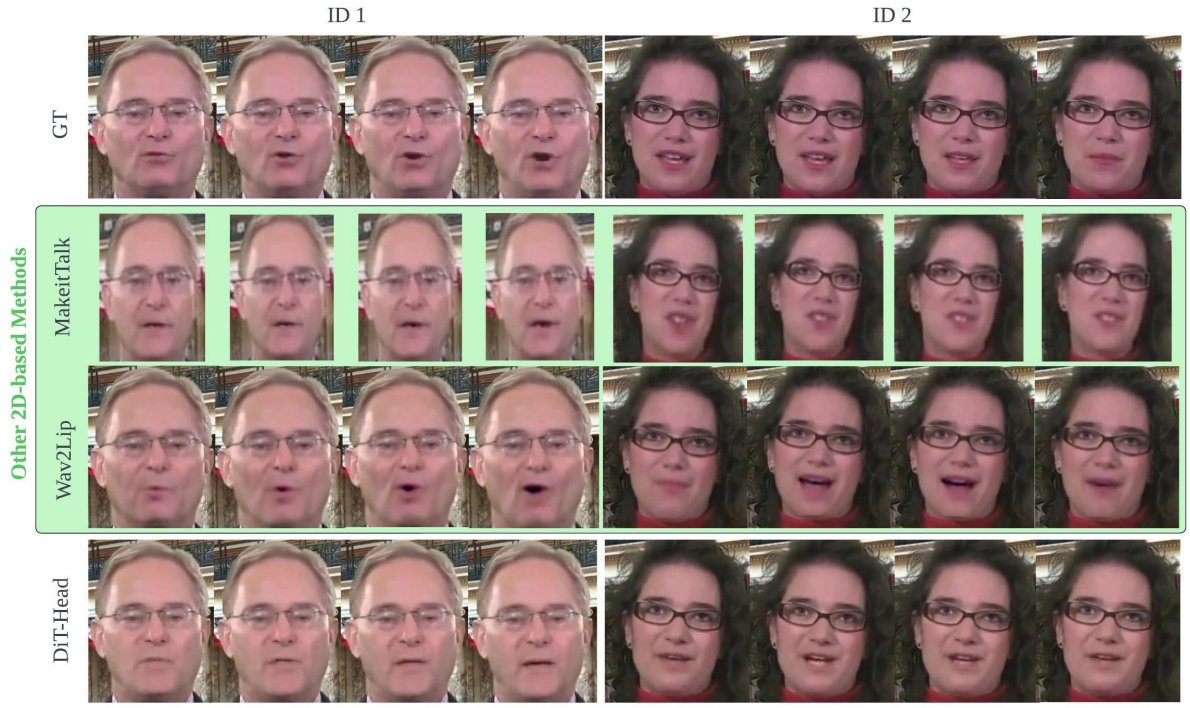

Figure 2: Qualitative comparison between DiT-Head and other 2D-based talking head synthesis methods (MakeItTalk (Zhou et al., 2020), Wav2Lip (Prajwal et al., 2020)). We encourage readers to view the supplementary video for a more accurate reflection of the qualitative differences between approaches.

¡blank line¿

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	LSE-C ↑	LSE-D ↓
MakeItTalk (Zhou et al., 2020)	22.47	0.73	0.207	55.20	3.40	11.4
Wav2Lip (Prajwal et al., 2020)	27.09	0.847	0.104	15.37	**10.78**	**5.32**
DiT-Head w/o Stage 3	**28.37**	**0.872**	**0.0856**	**10.31**	3.60	10.49

Table 1: Quantitative comparison on the test set between other 2D-based talking head synthesis methods (MakeItTalk (Zhou et al., 2020), Wav2Lip (Prajwal et al., 2020)) and DiT-Head w/o Stage 3. DiT-Head w/o Stage 3 was used for a fair comparison between frames. The best performance is highlighted in **bold**.

of quality, similarity and lip shape accuracy.

Table 1 suggests that DiT-Head achieves the highest PSNR (Kotevski and Mitrevski, 2010) and SSIM (Wang et al., 2004) compared to other methods, indicating the highest fidelity and similarity to the input video, as well as the lowest LPIPS (Zhang et al., 2018) and FID (Heusel et al., 2017), indicating the lowest perceptual and Fréchet distance from the input video. This suggests that DiT-Head preserves the identity and expression of the ground-truth better than the other methods. However, Wav2Lip (Prajwal et al., 2020) performs the best in both LSE-C (Prajwal et al., 2020) and LSE-D (Prajwal et al., 2020), indicating the most accurate lip shape on the center region and a similar lip shape on the whole region. This can be explained by the powerful lip-sync discriminator used in Wav2Lip (Prajwal et al., 2020) that ensures realistic lip movements and a perceptual loss that preserves the expression of the input. In contrast, DiT-Head outperforms MakeItTalk (Zhou et al., 2020) in both LSE-C (Prajwal et al., 2020) and LSE-D (Prajwal et al., 2020) and uses only a reconstruction loss with no lip-sync discriminator which trains in a more stable manner and is not prone to mode collapse.

Although we recognise lip-sync quality is highly subjective, the qualitative results in Figure 2 also highlight the benefits of our model over the existing methods. Our model can generate high-resolution outputs with smooth transitions between the generated and original regions, as it uses frame-interpolation (Huang et al., 2020) and a temporal audio window to enforce temporal context. This enhances the quality and sharpness of the face. Our model can also capture the pose and expression of

the input video better than MakeItTalk (Zhou et al., 2020), which uses a single identity image and audio to drive the talking head. Therefore, it cannot capture the identity and expression of the input as well as the other methods, nor can it generate realistic lip movements. Moreover, our model can avoid the blurriness and artifacts that affect Wav2Lip (Prajwal et al., 2020), which is trained on low-resolution images and generates a bounding box around the lower-half of the face. Both Wav2Lip (Prajwal et al., 2020) and DiT-Head produce less fine-detailed texture of the face when compared to the ground-truth which shows that both methods may lose some high-frequency information in the face.

5.1 Limitations and Future Work

Despite the promising results of our deep learning model for talking head synthesis, we acknowledge there are some limitations that need to be addressed in future work. A drawback of our work is that our model employs a ViT model (Dosovitskiy et al., 2020; Vaswani et al., 2017) that requires a lot of computational resources and training time to achieve its high performance for multi-modal learning. Furthermore, we only train using a pre-trained audio model (Baevski et al., 2020) for English speakers therefore DiT-Head cannot be effectively expanded to multiple languages.

Our model is scalable, but we aim to optimise our procedures to reduce the computational cost and speed up the training process by using flash-attention (Dao et al., 2022). We also plan to explore temporal finetuning of the autoencoders to address the temporal jitter issue (Blattmann et al., 2023). Moreover, we acknowledge that the inference process for diffusion models is slower than GAN-based approaches, which is an open research problem for diffusion models. However, we still achieve a speedup compared to most person-specific 3D-based methods. For example, DFRF (Shen et al., 2022) takes about 130 hours to finetune on a specific speaker and 4 hours to render a 55-second video at 720x1280 pixel size at 25 FPS (using default settings, excluding pre-processing or post-processing) on a single RTX 3090, while DiT-Head can produce a talking-head video of the same speaker with any driving audio in 8 hours on the same hardware. We believe that by addressing these limitations, our model can become a viable approach for person-agnostic talking head synthesis that can generate realistic and expressive videos.

5.2 Ethical Considerations

Talking head synthesis models can create realistic videos of people speaking with any content, but they can also create "deepfakes" that manipulate or deceive others (Korshunov and Marcel, 2022). Deepfakes can harm individuals and groups in various ways and erode the trust and credibility of information sources. Thus, methods to detect, prevent, and regulate the misuse of talking head synthesis models are needed. To address this threat, users of DiT-Head for talking head generation will need to authenticate their credentials and watermark generated videos. We did not watermark our training dataset due to time constraints, but we plan to do so in future work to avoid deepfakes generated by users of our model.

6 Conclusion

We have introduced a novel and advanced solution for talking head synthesis, DiT-Head, that does not depend on the person's identity and harnesses the powerful self- and cross-attention mechanisms of a DiT as opposed to UNets which suffer from a limited convolutional receptive field and GANs that suffer from unstable training and mode collapse. Our method outperforms existing ones in creating high-quality videos. The DiT enables our method to model the audio-visual-temporal dynamics of the input videos and produce realistic facial movements. In particular, the cross-attention mechanism allows our method to fuse audio and visual information for a more natural and coherent output. Our method opens up new possibilities for various applications in entertainment, education, and telecommunications.

REFERENCES

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Bigioi, D., Basak, S., Jordan, H., McDonnell, R., and Corcoran, P. (2023). Speech driven video editing via an audio-conditioned diffusion model. *arXiv preprint arXiv:2301.04474*.

Blanz, V. and Vetter, T. (1999). A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194.

Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., and Kreis, K. (2023). Align your latents: High-resolution video synthesis with latent diffusion models. *arXiv preprint arXiv:2304.08818*.

Bulat, A. and Tzimiropoulos, G. (2017). 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*.

Chen, L., Cui, G., Liu, C., Li, Z., Kou, Z., Xu, Y., and Xu, C. (2020). Talking-head generation with rhythmic head motion. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX*, pages 35–51. Springer.

Chen, L., Maddox, R. K., Duan, Z., and Xu, C. (2019). Hierarchical cross-modal talking face generationwith dynamic pixel-wise loss. *arXiv preprint arXiv:1905.03820*.

Cho, J., Lei, J., Tan, H., and Bansal, M. (2021). Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.

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

Cordonnier, J.-B., Loukas, A., and Jaggi, M. (2019). On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*.

Dao, T., Fu, D. Y., Ermon, S., Rudra, A., and Ré, C. (2022). FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*.

Dhariwal, P. and Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Doukas, M. C., Zafeiriou, S., and Sharmanska, V. (2021). Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14398–14407.

Esser, P., Rombach, R., and Ommer, B. (2021). Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883.

Geyer, C. J. (1992). Practical markov chain monte carlo. *Statistical science*, pages 473–483.

Guo, Y., Liu, Z., Chen, D., and Chen, Q. (2021). Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12388–12397.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

Huang, Z., Zhang, T., Heng, W., Shi, B., and Zhou, S. (2020). Rife: Real-time intermediate flow estimation for video frame interpolation. *arXiv preprint arXiv:2011.06294*.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Korshunov, P. and Marcel, S. (2022). The threat of deepfakes to computer and human visions. In *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*, pages 97–115. Springer International Publishing Cham.

Kotevski, Z. and Mitrevski, P. (2010). Experimental comparison of psnr and ssim metrics for video quality estimation. In *ICT Innovations 2009*, pages 357–366. Springer.

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., and Plumbley, M. D. (2023). Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*.

Lovelace, J., Kishore, V., Wan, C., Shekhtman, E., and Weinberger, K. (2022). Latent diffusion for language generation. *arXiv preprint arXiv:2212.09462*.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. (2021). Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Ngo, L. M., de Wiel, C. a., Karaoglu, S., and Gevers, T. (2020). Unified application of style transfer for face swapping and reenactment. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*.

Peebles, W. and Xie, S. (2022). Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*.

Prajwal, K., Mukhopadhyay, R., Namboodiri, V. P., and Jawahar, C. (2020). A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2021). High-resolution image synthesis with latent diffusion models.

Ronneberger, O., Fischer, P., and Brox, T. (2015). 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.

Shen, S., Li, W., Zhu, Z., Duan, Y., Zhou, J., and Lu, J. (2022). Learning dynamic facial radiance fields for few-shot talking head synthesis. In *European conference on computer vision*.

Shen, S., Zhao, W., Meng, Z., Li, W., Zhu, Z., Zhou, J., and Lu, J. (2023). Difftalk: Crafting diffusion models for generalized talking head synthesis. *arXiv preprint arXiv:2301.03786*.

Song, J., Meng, C., and Ermon, S. (2020). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Thome, N. and Wolf, C. (2023). Histoire des réseaux de neurones et du deep learning en traitement des signaux et des images. working paper or preprint.

Van Den Oord, A., Vinyals, O., et al. (2017). Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Y., Song, L., Wu, W., Qian, C., He, R., and Loy, C. C. (2022). Talking faces: Audio-to-video face generation. In *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*, pages 163–188. Springer International Publishing Cham.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Xu, C., Zhu, S., Zhu, J., Huang, T., Zhang, J., Tai, Y., and Liu, Y. (2023). Multimodal-driven talking face generation, face swapping, diffusion model. *arXiv preprint arXiv:2305.02594*.

Ye, Z., Jiang, Z., Ren, Y., Liu, J., He, J., and Zhao, Z. (2023). Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. *arXiv preprint arXiv:2301.13430*.

Yu, S., Sohn, K., Kim, S., and Shin, J. (2023). Video probabilistic diffusion models in projected latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zakharov, E., Shysheya, A., Burkov, E., and Lempitsky, V. (2019). Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9459–9468.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., and Li, S. Z. (2017). S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201.

Zhang, Y., He, W., Li, M., Tian, K., Zhang, Z., Cheng, J., Wang, Y., and Liao, J. (2022). Meta talk: Learning to data-efficiently generate audio-driven lip-synchronized talking face with high definition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4848–4852. IEEE.

Zhang, Z., Li, L., Ding, Y., and Fan, C. (2021). Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670.

Zhou, H., Liu, Y., Liu, Z., Luo, P., and Wang, X. (2019). Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9299–9306.

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