# READ Avatars: Realistic Emotion-controllable Audio Driven Avatars

Jack Saunders
https://jsaunders909.github.io/

Vinay Namboodiri
https://vinaypn.github.io/

Department of Computer Science
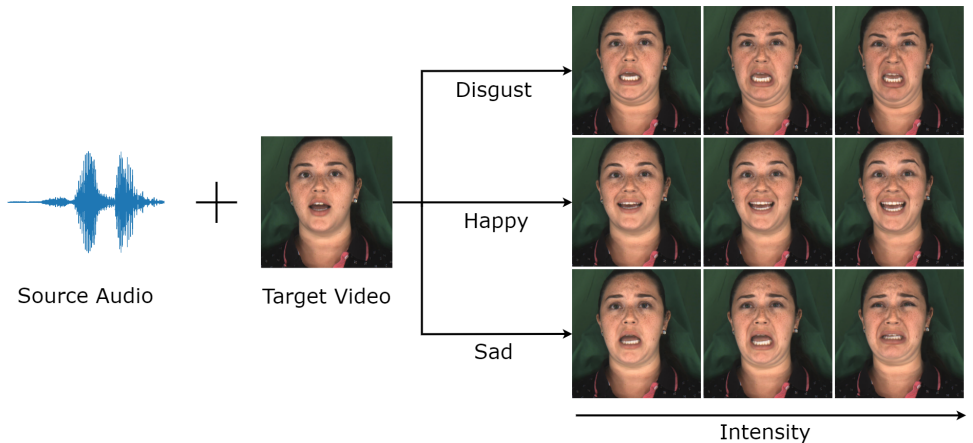University of Bath
Bath, UK

Figure 1: Our method, **READ Avatars** allows for the creation of photo-realistic and lip-synchronized video from audio and a reference video, with control over emotion. The same audio can be used to generate videos with multiple emotions.

## Abstract

We present READ Avatars, a 3D-based approach for generating 2D avatars that are driven by audio input with direct and granular control over the emotion. Previous methods are unable to achieve realistic animation due to the many-to-many nature of audio-to-expression mappings. We alleviate this issue by introducing an adversarial loss in the audio-to-expression generation process. This removes the smoothing effect of regression-based models and helps to improve the realism and expressiveness of the generated avatars. We note furthermore, that audio should be directly utilized when generating mouth interiors and that other 3D-based methods do not attempt this. We address this with audio-conditioned neural textures, which are resolution-independent. To evaluate the performance of our method, we perform quantitative and qualitative experiments, including a user study. We also propose a new metric for comparing how well an actor's emotion is reconstructed in the generated avatar. Our results show that our approach outperforms state-of-the-art audio-driven avatar generation methods across several metrics.

# 1   Introduction

Generating convincing talking head video is a highly desired capability in various applications, such as film and television dubbing, video games and photo-realistic video assistants. While significant progress has been made in this area [4, 10, 15, 17, 18, 23, 27, 30, 32, 33, 36, 38, 41, 44], most existing methods produce either low-quality but accurate lip sync using 2D models [3, 4, 7, 27, 32, 38] or high-quality but inconsistent lip sync using 3D models [17, 18, 23, 33, 36, 41]. We hypothesize that two key factors have prevented the development of models that are both high-quality and lip-synchronized. The first is that audio-to-expression is a many-to-many mapping. A given audio can correspond to many lip shapes, and the same lip shapes can produce different audio. The second factor is that while 3D models improve the visual quality by introducing strong priors, they struggle to represent complex lip shapes and do not model the mouth interior (see Figure 3).

Furthermore, it is desirable to introduce additional signals, such as emotion to offer users a greater level of control over the outputs. A few prior works attempt this [14, 15, 18, 26, 39]. These methods usually either consider emotion as discrete categories [18, 39], which gives semantic control but lacks granularity, or learn latent encodings of emotion [14, 15, 26] which allow for fine-grained control but is not semantic and requires selecting emotions from other sources (video or audio).

In this paper, we introduce READ Avatars, a method for generating talking head videos with direct and granular control over emotion, while achieving high levels of lip sync, emotional clarity, and visual quality. We build upon 3D-based approaches, using a morphable model [22] as an intermediate representation of the face and deferred neural rendering [36] to achieve high visual quality. To address the above issues causing poor lip sync in 3D models, we propose two novel components. First, we add an adversarial loss to the audio-to-expression generator to alleviate the many-to-many mapping issue. Second, we overcome the challenge of representing complex lip shapes and mouth interiors with a morphable model by conditioning a neural texture on audio, encoding audio features on the surface of the mesh using a resolution-independent neural texture based on a SIREN network [31].

In summary, our contributions are:

- **A novel neural rendering algorithm** that leverages neural textures, operates directly on UV coordinates, and can be conditioned on audio, improving the mouth interior.

- **The incorporation of a GAN loss into the audio-to-expression network** to improve the results by solving the many-to-many issue of audio-to-expression generation.

- **A new metric** for determining how well an actor's emotions are captured and reconstructed.

# 2   Related Works

## 2.1   Unconditional Audio-Driven Face Models

**2D Models:** Many approaches to synthesizing talking head videos from audio operate directly in the image or video domain [4, 27, 32]. These methods typically employ an encoder-decoder architecture. ATVG [4] uses audio to control 2D landmarks, which are then used
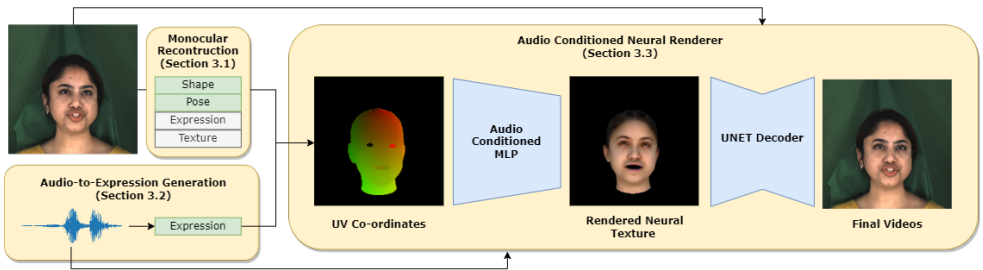
Figure 2: The READ Avatars pipeline. We train 3 separate networks: an audio-to-expression network (Section 3.2), an audio-conditioned MLP-based network that converts the UV coordinates to neural texture features (Section 3.3), and a final UNET decoder network.

to generate video with attention to highlight the parts that need editing. Wav2Lip [27] significantly improves lip sync accuracy by minimizing the distance between the audio and generated video according to a pre-trained lip sync detection network. While the lip sync is excellent, the visual quality is poor. Recently, a context-aware transformer [32] was applied to this problem, with an audio-injected refinement network that significantly improves the visual quality. However, all 2D-based models to date suffer from limited visual quality. In contrast, our 3D-based method produces much higher-quality videos.

**3D Guided Face Models:** Using explicit 3D supervision, ultra high-quality face models driven by various signals have been created [17, 18, 20, 23, 34, 36, 41]. These methods simplify facial synthesis by modeling the underlying 3D scene with a small set of parameters, such as a 3DMM [2, 6], that can be directly controlled. Despite their high visual quality, these models often lack expressiveness due to the many-to-many mapping problem and the limited lip expressions of the underlying geometry. Puppetry methods [17, 18], and motion models [15, 30] are able to somewhat solve the many-to-many issue by using a source actor to drive the expressions. This provides a signal, the source actor's expressions, which is much closer to one-to-one with the target actor. However, it is often undesirable to require a source actor to be filmed, and the resulting video processed every time the model is used. Actor-free methods such as ours are significantly more scalable. Implicit models [9, 10, 25, 45] augment geometry using MLP offsets, allowing for more expressive lip shapes but do not solve the many-to-many problem. Concurrent work [33] addresses both the many-to-many issue and the mouth interior using memory networks. However, they rely on reusing explicit pixels from the mouth region, which leads to jitter in the final videos.

## 2.2 Audio-Driven Face Models with Emotional Control

Only a small number of works have attempted to develop models that allow for explicit control of stylistic attributes, such as emotion, in generated talking head videos. MEAD [39] introduces an audio-driven model with control over emotion. They trained a network to map audio to landmarks and another to convert input images to the desired emotion and then used a UNET-based network to combine the upper face with the desired emotion and the generated landmarks to produce the final image. This method can control emotion and intensity, but lacks temporal coherence due to its frame-by-frame nature and has suboptimal lip sync. EVP [14] improves upon MEAD by adding an emotion disentanglement network to separate content and emotion in the audio and using a face-synthesis network [40] to produce

higher-quality videos with temporal consistency. However, the lip sync and emotional clarity were not always satisfactory. MEAD and EVP are the most similar to our work. These landmark-based models are unable to produce as high-quality results as 3D models, they also suffer from unnatural motion without the strong priors of a 3DMM. Some methods are able to overcome the many-to-many problem by using a source actors expressions [15, 18] or a person-generic model [26]. These methods produce high-quality results but require significant processing for each actor. Furthermore, the mismatch in expressions between the source and target actor can create artefacts.

## 3 Method

Our method consists of three stages. In the first stage (Section 3.1), we fit a 3D Morphable Model [0, 21, 22, 46] to the input videos using an extension of the Face2Face monocular reconstruction algorithm [9, 54]. In the second stage (Section 3.2), we train a neural network inspired by Pix2Pix [13] to generate morphable model parameters from audio using adversarial training. This allows us to generate realistic animation sequences, even in areas that are not well correlated with audio. At this stage, we introduce a fine-grained emotional label. In the final stage (Section 3.3), we train an audio-conditioned deferred neural renderer [55]. Our model uses a SIREN MLP [51] to directly map uv coordinates to texture features, replacing the learned neural texture of previous work [55], and making the task of audio conditioning much easier.

### 3.1 Monocular Reconstruction

In the first stage of our method, we aim to find a low-dimensional set of parameters that can model a video sequence $V = (V_0, \ldots V_n)$. For this purpose, we use the FLAME model [22]. The FLAME model can be represented as a function $\mathcal{V}$ that maps a set of parameters for shape $\beta$, the expression $\psi$, and joint rotations $\theta$ onto 5023 3D vertices. A similar function is used to map a set of texture parameters $\alpha$ onto UV-based 2D textures. We model the rendering process using a full perspective camera that projects a mesh onto the image plane with parameters $\mathbf{R}$ and $\mathbf{t}$, the rotation and translation and $K$, the camera intrinsics. We model lighting using spherical harmonics with parameters $\gamma$. If we define $\pi$ as the set of all the above parameters, then our objective is to find the optimal $\tilde{\pi}$ that best fits a given image. To fit the FLAME model to our videos, we adopt the tracking model of Neural Head Avatars [9], which is based on Face2Face [54]. This model uses differentiable rendering in Pytorch3D [28] to minimize the $L_1$ distance between real and rendered frames, with statistical regularization over the parameters.

We assume that shape, texture, and lighting are fixed for a given actor, as our data is captured under controlled conditions. Therefore, we can first estimate $\pi_{fix} = (\alpha, \gamma, K)$, the parameters that are fixed across all videos for a given subject. These parameters are then fixed for all frames in all videos of the same subject. We can then estimate $\pi_{var} = (\theta, \psi, \mathbf{R}, \mathbf{t})$ on a per-frame basis. These parameters are then the target of the audio-to-parameter generator.

### 3.2 Audio-to-Parameter Generator

The goal of our method is to animate photo-realistic avatars using audio as the control signal. Previous approaches based on 3DMM's [18, 20, 36, 41] use audio-to-parameter models

which rely on neural networks with regression losses to generate a subset of the parameters. Such methods, however, suffer from the many-to-many issue when mapping audio to expressions. Regression-based losses mean that weakly correlated parameters such as those involved in upper face motion are almost entirely averaged out, while even highly correlated parameters such as the lip and jaw are over-smoothed. To address these issues, we propose an audio-to-parameter generator based on a conditional GAN [8]. Our model is similar to Pix2Pix [13], using a combination of $L_1$ loss for low-frequency parts of the data and an adversarial loss for increased realism.

The input of this network is a section of audio represented as MFCC coefficients, $\mathbf{A}$, together with an explicit emotion label. The window size of the MFCC is selected to be a multiple of the video frame rate. We use an explicit emotion labelling system to introduce emotion into the generated parameters. For $N$ emotions, we use an $N-1$ dimensional label, $C$, with neutral emotion represented as a zero vector (absence of emotion). Each other emotion is assigned to a dimension and scaled by intensity, with the maximum intensity being 1. This continuous label allows for fine-grained control over the emotion. We distribute the label over the time dimension to obtain $\mathbf{C} = (C_0, \dots, C_n)$. This label is concatenated with the MFCC audio features and serves as the input to the audio to expression generator $\mathcal{G}_a$, which produces the target parameters for each frame $(\pi_0, \dots \pi_n)$.

The discriminator is conditional, and takes either the real or generated parameters, together with the audio and emotional label and predicts if the given parameters are real or generated. Both the generator and discriminator networks use an encoder-temporal-decoder model, projecting the audio features into a high-dimensional latent space via a fully connected layer followed by an LSTM [12] and a fully connected decoder to map from this latent space to the parameters. More details can be found in the supplementary material. We optimize the objective:
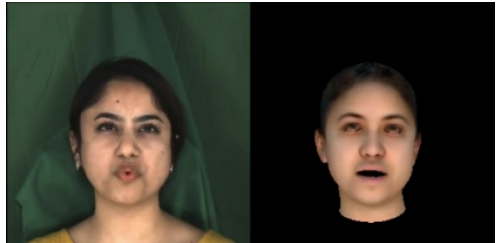


Figure 3: An example of a typical failure in the monocular reconstruction method. *(Left):* The input frame, *(right):* the reconstruction. Note the failure to capture the "O" shape.

$$\mathcal{L} = \mathcal{L}_1 + \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} \tag{1}$$

where $\mathcal{L}_1$ is the $\ell_1$ distance between the real and predicted parameters, $\mathcal{L}_{\text{GAN}}$ is the adversarial loss and $\mathcal{L}_{\text{vel}}$ is an $\ell_1$ distance between the velocities of the output animation. The velocity loss is known to improve the temporal consistency of speech-driven animation [5]. Each $\lambda$ is a relative weight, we use $\lambda_{\text{GAN}} = 0.02$ and $\lambda_{\text{vel}} = 100$.

## 3.3 Audio-Conditioned Neural Renderer

We next consider how to invert the parametric model fitting and produce photo-realistic video given a set of parameters $(\pi_i)$. We build upon the idea of neural textures [55], jointly optimizing an image-to-image deferred neural renderer, and a neural texture defined in UV space. However, we find that the FLAME model is not expressive enough to represent complex lip motions (see Figure 3), and that neural textures alone are not sufficient to compensate for

this. Furthermore, the FLAME mesh provides no information about the interior of the mouth including the tongue and teeth.

To address this issue, we propose audio-conditioned neural textures. The aim is to encode audio information on the surface of the mesh to allow for more complex lip shapes and mouth interiors to be learned. We replace the static neural texture with a SIREN MLP [31] texture network $\mathcal{T}$, which maps a uv coordinate of the rasterized meshes directly to a feature vector.

This bypasses the need for texture lookup which is slow and limits the resolution of the neural textures. The use of a network also allows us to easily condition on audio by simply concatenating audio features to the UV coordinates. It is important that the audio features used in the network do not contain emotion or identity information. To remove such information, we use the output of a Wav2Vec2 network [1, 42] pretrained to predict phoneme probabilities at a 50 fps. The 50 most common phonemes comprise over 99% of the audio data, therefore we restrict the features to these only. Next, we resample these probabilities to 60 fps, twice the frame rate of the video. We take a window of $W$ frames centred at the target video frame and use a neural network $\mathcal{A} : R^{2W \times 50}$ to encode the audio over this window into a single vector $\mathbf{a}_{enc} \in R^{N_a}$, where $N_a$ is the dimension of the encoded audio and is a hyperparameter. The encoded audio vector is then used to condition the neural texture. The audio encoder makes use of 1D temporal convolutions. The now encoded audio vector $\mathbf{a}_{enc}$ is concatenated with the UV coordinates obtained during rasterization, which serves as the input to our texture network. The output of the texture network is a multi-channel image, which appears as a rasterization of the mesh with an audio-encoded neural texture. Similar to [35], we use a 16-channel neural texture, enabling the representation of higher-order lighting effects. This rasterized image is then passed through a UNET [29] based deferred neural renderer $\mathcal{R}$, which produces a photorealistic final frame, leveraging the audio features encoded on the mesh. Details of the SIREN texture network and the UNET can be found in the supplementary material.

To address the issue of jitter in the final video, we include additional renderings for the frames in a window of length $W_R$ centred on the target frame. These additional renderings are based on the rasterized UV coordinates of the parameters from the frames on either side of the target frame. This results in an input with $16 + 2W_R$ channels for the UNET decoder, which is able to smooth out the jitter over a window of frames.
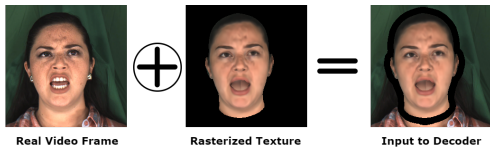


Figure 4: The input to the decoder network consists of a rendered neural texture and a real frame with a border of black pixels. Note that the texture cannot represent the mouth interior, but it is generated by the decoder.

To produce realistic and temporally stable background in our videos, we blend the output of our texture network $\hat{V}_t$ with the original video frames $V_t$. We use the alpha channel from the rendered mesh as a mask to separate the foreground and background. The foreground mask, $\alpha$ is expanded by a fixed number of pixels to obtain $\alpha_{exp}$. This expanded mask is zeroed out of the real frame and the foreground mask $\alpha$ is used to fill in these pixels with the rendered mesh. This is best shown in Figure 4.

$$V_t^* = \alpha \hat{V}_t + (1 - \alpha_{exp}) V_t \qquad (2)$$

To train the audio encoder, texture network, and deferred neural renderer, we optimize the following objective function end-to-end:
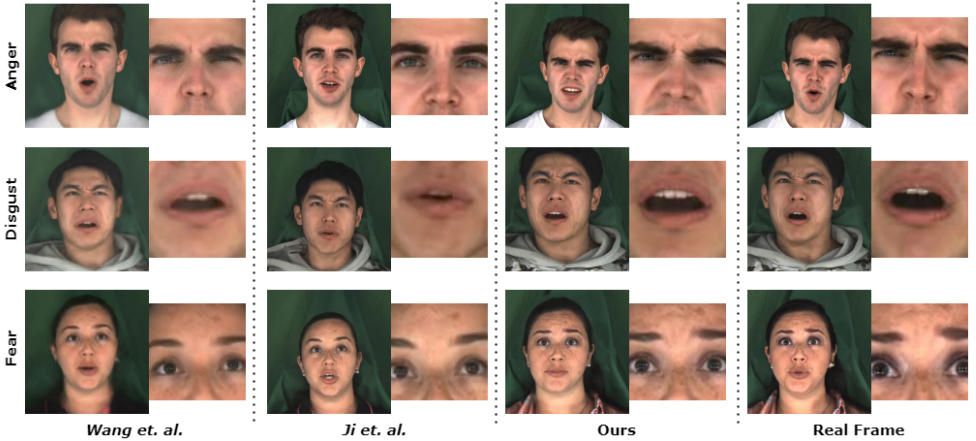
Figure 5: We compare our results to those of MEAD [39] and EVP [14]. Our results are of much higher visual quality than those of MEAD, the zoomed-in regions demonstrate that our method produces more convincing and accurate emotions compared to EVP.

$$\mathcal{L}(V^*, V) = \lambda_1 \mathcal{L}_1 + \lambda_{\text{VGG}} \mathcal{L}_{\text{VGG}} + \lambda_{\text{GAN}} \mathcal{L}_{\text{GAN}} \tag{3}$$

Here, $\mathcal{L}_1$ is the $\ell_1$ distance between the real and generated frames, $\mathcal{L}_{VGG}$ is a VGG-based style loss [16], and $\mathcal{L}_{\text{GAN}}$ is an adversarial loss. The hyperparameters $\lambda_1$, $\lambda_{\text{VGG}}$, and $\lambda_{\text{GAN}}$ are used to weight the importance of each loss. We use $\lambda_1 = \lambda_{\text{VGG}} = 1$ and $\lambda_{\text{GAN}} = 0.01$.

## 3.4 Implementation Details

We first crop every frame to a square shape of 256 pixels. We do this by estimating a bounding box for each frame with padding and then finding the smallest square that covers the union of these boxes. We reshape this square to the desired resolution. We implement our pipeline in Pytorch and Pytorch3D. All networks are optimized using Adam [19] with a learning rate of 0.0001. The renderer is trained for 5 epochs with a batch size of 1, while the audio-to-expression generator is trained for 10 epochs with a batch size of 32. We use the LSGAN formulation for all adversarial training [24]. Further details may be found in the supplementary material.

# 4 Results

We use the MEAD dataset [39] for our experiments, which includes 60 actors (both male and female) speaking 30 sentences in 8 different emotions at 3 levels of intensity. We follow the train-test split outlined in MEAD and train models for 4 of these subjects.

In order to evaluate the performance of our method, we consider three qualities: visual quality, lip sync, and emotional clarity. For visual quality, we use the Fréchet Inception Distance (FID) metric. To measure lip synchronization, we use the Lip Sync Error (LSE) metrics introduced in wav2lip [27]. These metrics are calculated using a pre-trained syncnet and include LSE-D, which measures the minimum distance between audio and video features, and LSE-C, which measures the confidence that the audio and video are synchronized.

Table 1: Results of the user study across 4 categories. Where ours is preferred strongly (weakly) we denote the result $++$ $(+)$, where there is no preference, 0 and where the other method is preferred strongly (weakly), $--$ $(-)$.

| Statement | $--$ | $-$ | 0 | $+$ | $++$ | mean |
|---|---|---|---|---|---|---|
| Ours > MEAD (lip-sync ) | 1 | 12 | 14 | 28 | 44 | $+1.02$ |
| Ours > MEAD (visual quality) | 0 | 3 | 9 | 23 | 65 | $+1.49$ |
| Ours > MEAD (naturalness) | 1 | 1 | 15 | 23 | 65 | $+1.39$ |
| Ours > MEAD (emotion) | 1 | 3 | 21 | 36 | 38 | $+1.06$ |
| Ours > EVP (lip-sync ) | 7 | 16 | 22 | 42 | 11 | $+0.50$ |
| Ours > EVP (visual quality) | 2 | 22 | 23 | 42 | 11 | $+0.39$ |
| Ours > EVP (naturalness) | 7 | 17 | 18 | 26 | 38 | $+0.49$ |
| Ours > EVP (emotion) | 4 | 8 | 22 | 26 | 38 | $+0.87$ |
| Ours > Real (lip-sync ) | 69 | 28 | 3 | 1 | 0 | $-1.64$ |
| Ours > Real (visual quality) | 33 | 41 | 23 | 2 | 0 | $-1.05$ |
| Ours > Real (naturalness) | 53 | 38 | 8 | 1 | 0 | $-1.43$ |
| Ours > Real (emotion) | 40 | 38 | 19 | 3 | 1 | $-1.13$ |

To measure emotional clarity, it is not enough to measure differences at the frame level, as the intensity of emotion naturally varies over a video. We, therefore, introduce a new metric for emotional clarity that measures the differences between distributions of emotion. This metric is based on a pre-trained EmoNet model [57]. We predict the valence and arousal for each frame and compare the distributions of these values between the generated and ground truth videos. We approximate the distance between these distributions using the Earth Movers Distance to obtain the valence Emotional Mean Distance (V-EMD) and arousal Emotional Mean Distance (A-EMD) metrics.

## 4.1 Comparisons to State-of-the-Art



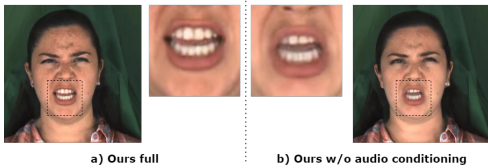a) Ours full    b) Ours w/o audio conditioning

Figure 6: The addition of audio conditioning in the neural texture improves the quality of the resulting frame, particularly in the mouth interior.

We compare our method to several state-of-the-art audio-driven avatar models. Our main comparisons are with Audio-Driven Expressive Video Generation (EVP) [14] and Multimodal Emotion-aware Dataset (MEAD) [39]. MEAD uses an audio-to-landmark LSTM, an emotion transformer to alter the audio-driven landmarks to any given emotion, and a final UNET-based model to produce output frames from the emotional landmarks. AudioDrivenEVP improves on this approach by designing a disentanglement model to separate audio into emotion and content, which is then used with a landmark alignment method to control for pose and a video-to-video network that produces high-quality and temporally stable video from landmarks. We also compare to ATVG [4], a 2D-based method that excels in lip synchronization but has poor visual quality, and is unable to edit emotion.

**Quantitative:** The results of these comparisons are shown in Table 2. Our results outperform all competitors on visual quality (FID) and emotional reconstruction (A/V-EMD).

Table 2: **Quantitative comparisons to state-of-the-art**. We compare visual quality using FID, lip sync with LSE-D/C [27] and emotional reconstruction with our metrics A/V-EMD. We compare our results with ATVG [4], MEAD [39] and Audio-driven Emotional Video Portraits [14] (EVP)

| Method | LSE-C ↑ | LSE-D ↓ | FID ↓ | A-EMD ↓ | V-EMD ↓ |
|--------|---------|---------|-------|---------|---------|
| ATVG [4] | **5.705** | **8.731** | 120.040 | 0.160 | 0.239 |
| MEAD [39] | 4.080 | 10.569 | 38.015 | 0.974 | 0.113 |
| EVP [14] | 4.061 | 11.514 | 43.972 | 0.119 | 0.126 |
| Ours | 4.431 | 10.157 | **13.600** | **0.0686** | **0.093** |

Table 3: **Ablation study**. We compare visual quality using FID, lip sync with LSE-D/C [27] and emotional reconstruction with our metrics A/V-EMD. We compare our full model to the same model both without the GAN loss in the audio-to-expression generator and without the audio-conditioned neural texture.

| Method | LSE-C ↑ | LSE-D ↓ | FID ↓ | A-EMD ↓ | V-EMD ↓ |
|--------|---------|---------|-------|---------|---------|
| Ours | **4.431** | **10.157** | 13.600 | **0.069** | 0.093 |
| Ours w/o GAN loss | 4.047 | 10.446 | **12.587** | 0.079 | **0.090** |
| Ours w/o audio texture | 4.175 | 10.398 | 15.96 | 0.069 | 0.96 |

While our method is not able to reach the lip-sync quality of the unconditional ATVG [4], it has far better visual quality and emotional clarity. We outperform both methods capable of controlling emotion: MEAD [39] and EVP [14] on lip sync. **Qualitative:** Figure 5 shows our results in comparison to MEAD [39] and EVP [14]. Our results show clearly better visual quality than MEAD. Compared with EVP our method is capable of producing emotion that is much clearer and more closely matches the real videos, the expanded regions highlight this. In particular, it can be seen that the eyebrows convey the target emotion far better in our method. Additional results can be found in the supplementary video that further demonstrates the advantages of our method. **User Study:** To gauge the subjective quality of our generated avatars, we conducted a user study. We selected four subjects and generated five videos in each of the eight emotions for a total of 160 videos. We perform a two-alternative forced choice study, pairing each of our videos with its counterpart from the alternatives. The users are shown both of the videos in a random order, together with the target emotion. We ask users to select which of the two videos is better in four categories: lip sync, visual quality, naturalness and emotional clarity. A total of 10 users completed the study. The results are shown in Table 1. Our method strongly outperforms MEAD across all categories. Compared with EVP, our method is also preferred across all categories. However, this preference is weaker for lip sync and natural quality, but much stronger for emotional clarity.

## 4.2 Ablation Study

We also perform an ablation study for both the adversarial loss and the audio-conditioned neural texture. The results of this comparison are shown in Table 3. The inclusion of the adversarial loss improves the lip-sync at the cost of a small loss in visual quality. For the audio-conditioned neural texture, we compare our work to a static neural texture [35]. Our method improves the visual quality and lip-sync, with small improvements in the emotional reconstruction. As expected, the improvements resulting from our audio conditioning are

most notable in the mouth interior. This is because the audio allows the decoder to disambiguate the multiple mouth interiors that could be represented by the same underlying geometry. Figure 6 shows this improvement. These results are further enforced by a further user study. This time, we ask users only to select their preference between the full and ablated methods. The results are shown in the supplementary material and show that our novel components do indeed improve the results, as judged by users.

# 5  Conclusion

We present a new method for producing audio-driven avatars with control over emotion. We have used a 3D-based pipeline with the novel additions of an adversarial loss in the audio-to-expression generator and an audio-conditioned, resolution-independent neural texture. Our method alleviates the many-to-many problem in conditioned, audio-driven video generation and surpasses state-of-the-art for lip sync and visual quality, as well as emotional reconstruction, as highlighted by our novel metric. Our comprehensive solution can be used for diverse applications. **Limitations:** Our model sometimes suffers when using extreme poses (Figure 7). Fur-
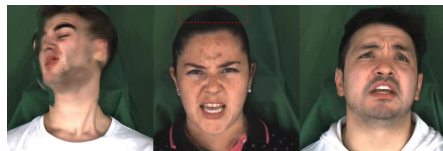


Figure 7: Failure cases of our method. **Left & middle:** When the pose of the target video is significantly different from the training data, artefacts occur. **Right:** When the tracking is inaccurate, our model produces blurry results.

thermore, as we use reference videos to control for pose and background, the length of the generated videos is limited. Future work will look to address arbitrary length video generation, potentially by considering pose generation. It is also worth investigating other models that address many-to-many generation, such as diffusion models [11, 43]. **Ethical Implications:** The ability to create synthetic digital humans comes with a serious potential for misuse. For this reason, we do not make our pipeline available to the general public. However, we are willing to share code with other researchers. As a person-specific model, we will follow ethical guidelines to require permission from any user to create their model.

# 6  Acknowledgments

# References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

[2] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, page 187–194, USA, 1999. ACM Press/Addison-Wesley Publishing Co. ISBN 0201485605. doi: 10.1145/311535.311556. URL https://doi.org/10.1145/311535.311556.

[3] 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*, SIGGRAPH '97, page 353–360, USA, 1997. ACM Press/Addison-Wesley Publishing Co. ISBN 0897918967. doi: 10.1145/258734. 258880. URL https://doi.org/10.1145/258734.258880.

[4] Lele Chen, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[5] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael Black. Capture, learning, and synthesis of 3D speaking styles. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10101–10111, 2019.

[6] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3d morphable face models—past, present, and future. *ACM Trans. Graph.*, 2020.

[7] Tony Ezzat, Gadi Geiger, and Tomaso Poggio. Trainable videorealistic speech animation. *ACM Trans. Graph.*, 21(3):388–398, jul 2002. ISSN 0730-0301. doi: 10.1145/ 566654.566594. URL https://doi.org/10.1145/566654.566594.

[8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[9] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. *arXiv preprint arXiv:2112.01554*, 2021.

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

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2016.

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

[15] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393379. doi: 10.1145/3528233.3530745. URL https://doi.org/10.1145/3528233.3530745.

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[17] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):163, 2018.

[18] Hyeongwoo Kim, Mohamed Elgharib, Hans-Peter Zollöfer, Michael Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt. Neural style-preserving visual dubbing. *ACM Transactions on Graphics (TOG)*, 38(6):178:1–13, 2019.

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

[20] 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 (CVPR)*, 2021.

[21] J. P. Lewis, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Fred Pighin, and Zhigang Deng. Practice and Theory of Blendshape Facial Models. In Sylvain Lefebvre and Michela Spagnuolo, editors, *Eurographics 2014 - State of the Art Reports*. The Eurographics Association, 2014. doi: 10.2312/egst.20141042.

[22] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. URL https://doi.org/10.1145/3130800.3130813.

[23] Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live speech portraits: Real-time photorealistic talking-head animation. *ACM Trans. Graph.*, 40(6), dec 2021. ISSN 0730-0301. doi: 10.1145/3478513.3480484. URL https://doi.org/10.1145/3478513.3480484.

[24] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[25] 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. In *ECCV*, 2020.

[26] Foivos Paraperas Papantoniou, Panagiotis P. Filntisis, Petros Maragos, and Anastasios Roussos. Neural emotion director: Speech-preserving semantic control of facial expressions in "in-the-wild" videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[27] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. 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*, MM '20, page 484–492, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413532. URL https://doi.org/10.1145/3394171.3413532.

[28] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

[30] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, December 2019.

[31] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Proc. NeurIPS*, 2020.

[32] 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*, SA '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394703. doi: 10.1145/3550469.3555393. URL https://doi.org/10.1145/3550469.3555393.

[33] Anni Tang, Tianyu He, Xu Tan, Jun Ling, Runnan Li, Sheng Zhao, Li Song, and Jiang Bian. Memories are one-to-many mapping alleviators in talking face generation, 2022. URL https://arxiv.org/abs/2212.05005.

[34] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. *Commun. ACM*, 62(1):96–104, dec 2018. ISSN 0001-0782. doi: 10.1145/3292039. URL https://doi.org/10.1145/3292039.

[35] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4): 1–12, 2019.

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

[37] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 2021. URL https://www.nature.com/articles/s42256-020-00280-0.

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

[39] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020.

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

[41] Xin Wen, Miao Wang, Christian Richardt, Ze-Yin Chen, and Shi-Min Hu. Photorealistic audio-driven video portraits. *IEEE Transactions on Visualization and Computer Graphics*, 26(12):3457–3466, 2020.

[42] Qiantong Xu, Alexei Baevski, and Michael Auli. Simple and effective zero-shot cross-lingual phoneme recognition. *arXiv preprint arXiv:2109.11680*, 2021.

[43] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*, 2022.

[44] Xinwei Yao, Ohad Fried, Kayvon Fatahalian, and Maneesh Agrawala. Iterative text-based editing of talking-heads using neural retargeting. *ACM Trans. Graph.*, 40(3), aug 2021. ISSN 0730-0301. doi: 10.1145/3449063. URL https://doi.org/10.1145/3449063.

[45] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. I M Avatar: Implicit morphable head avatars from videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2022.

[46] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, Thabo Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. *Computer Graphics Forum*, 37:523–550, 05 2018. doi: 10.1111/cgf.13382.