Comparing Vision Generative Models on Talking-Head Synthesis

Songyu Han
Institute for Computational
& Mathematical Engineering
Stanford University
songyuh@stanford.edu

Mustafa Abdelrahim Haroun Fadl
Institute for Computational
& Mathematical Engineering
Stanford University
mustafaf@stanford.edu

Hannah Levin
Computer Science
Stanford University
levinh@stanford.edu

Abstract

In this work, we investigate the effectiveness of Generative Adversarial Networks (GANs) and Denoising Diffusion Probabilistic Models (DDPMs) for the task of synchronized talking-head video synthesis using only an identity frame and an audio clip as inputs. Our goal is to generate short video sequences with realistic faces while minimizing additional supervision. We implemented both approaches and evaluated them in the CREMA-D dataset. Our DDPM achieved significantly better results, with an FID of 81.68, which markedly outperformed GANs, which recorded an FID of 94.26. These findings underscore the robustness of diffusion-based models in generating high-fidelity, audio-synchronized facial animation with minimal supervision.

1. Introduction

Animating talking-heads is an important task in filmmaking, game development, and virtual reality applications, and it is traditionally seen as a task that relies on substantial manual input. However, with the recent advancements in generative computer vision, producing high-fidelity video clips featuring talking heads became a key research topic in the area of conditional generative models. Traditionally, producing animated clips of talking-heads that are consistent with human perception is a difficult task due to the following three challenges: First, the high-dimensional nature of facial dynamics [20] makes consistent landmark generation difficult, second, many models struggle to reproduce natural motions beyond lip movement, such as blinking and head movements, and third, most state-of-the-art methods often require additional supervision to supplement the audio signals in order to generate naturally-looking clips.

In this project, we aim to study the relative strengths of Generative Adversarial Networks (GANs) and Diffusion Models on the task of synchronized talking-head synthesis. Specifically, our goal is to utilize audio containing human speech and a reference identity frame of the speaker as inputs to our models and produce short video clips of talking-heads while using minimal additional supervision.

Our experimental results demonstrate a clear distinction between the capabilities of GANs and diffusion models for this task. While GANs achieved basic lip synchronization, they often suffered from training instability, limited visual fidelity, and an inability to reproduce natural head or eye movements. In contrast, diffusion models—particularly those enhanced with facial landmark conditioning—produced temporally coherent sequences with sharper visuals and more accurate synchronization. Quantitatively, as will be detailed, the diffusion model significantly outperformed the GAN across key metrics for visual quality (FID 81.68 vs. 94.25), audiovisual synchronization (AV offset: 0.16 vs. –3.48), and perceptual confidence (AV confidence: 2.44 vs. 2.68). These findings strongly suggest that diffusion-based architectures provide a more effective pathway for this domain.

2. Related Works

The problem of talking head video synthesis can be roughly grouped into two categories: speech-driven synthesis and motion-driven synthesis, and we are mostly interested in the former. Early attempts at capturing correlations between speech and facial features in computer graphics were mostly centered around Hidden Markov Models (HMMs), with Brand's Voice Puppetry [2] as well as Xie's Coupled HMM [18] being notable pioneers in the field. However, the advent of Generative Adversarial Networks architectures [6] has driven the most remarkable progress in image and video synthesis, and notable speech-driven

talking-head video generation models such as Prajwal et. al.'s Wav2Lip [10] and Zhou et. al.'s MakeItTalk [20] all use the GAN architecture.

However, even though GANs were able to achieve the state-of-the-art in terms of generation fidelity, past research have found numerous challenges in GAN training. First, GANs are optimized using a minimax objective, which is liable to a phenomenon known as "convergence failure" when the discriminator starts to dominate the generator or vice versa. [1] Second, GANs are also liable to "mode collapse" (a phenomenon referring to the model's inability to generate a wide variety of samples) if the hyperparameters and regularizers are not chosen carefully. [5]. To address GAN's training instability problems, in 2020, researchers at UC Berkeley proposed a new class of generative models based on probabilistic Markov chains named "Denoising Diffusion Probabilistic Models" (DDPMs). [7] Even though DDPMs are more computationally inefficient at processing smaller datasets, more difficult to implement, and slower at inference as compared to GANs, DDPMs are still favored over GANs in most generative vision tasks thanks to their training stability and scalability. [9]

More recent progressions in the field of speech-driven talking head video synthesis mostly incorporate DDPMs as the backbone, with examples being Shen et. al.'s DiffTalk [12] and Stypulkowski et. al.'s Diffused Heads [14]. Our project directly builds on top of the ideas presented by Shen et. al. and Stypulkowski et. al. and adapts their model architectures to settings where training data and computation resources may be limited.

3. Datasets

We used the CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset for our project. This dataset contains 7442 short video clips of 91 actors with diverse ethnic backgrounds speaking short English sentences in a wide range of emotional states. All video clips in this dataset consist of an actor speaking in front of a green screen background. [3]

The original video clips are recorded in a resolution of 360 x 480 and a frame rate of 29.97. A significant problem we found with these original clips is that the relative position of the talking head in the frame varies significantly due to the actors' hight differences, so we were able to source a processed version of CREMA-D with the videos preprocessed to have the talking heads centered within each cropped 320 x 320 frame. We also adjusted the video framerate to 25 and the audio sample rate to 16,000 so that the timespan of one frame matches with the timespan of one audio embedding. Since extracting audio features is not a focus of our project, we simply used the pre-trained audio encoder used by Stypulkowski et. al. and Vougioukas et. al. [16]

During training, we downsampled all video frames to either 64 x 64 or 128 x 128 in order to reduce computational costs. We also extracted facial landmarks from each frame using the DLib face recognition toolkit. We performed a 90-10 Train-Validation split on our dataset on the **actor** level, meaning the validation set is entirely comprised of clips with faces that our models have not seen during training.

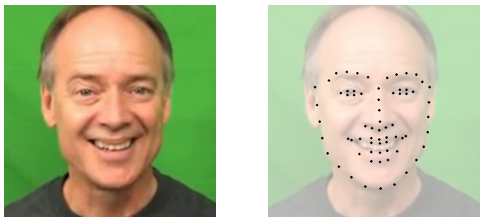


Figure 1: Sample of video frame and corresponding landmarks

4. Methods

We implemented two different model architectures for synchronized talking-head synthesis: a Generative Adversarial Network (GAN) and a Denoising Diffusion Probabilistic Model (DDPM). Specifics of model architectures, training procedures and hyperparameters will be discussed in detail below.

4.1. Generative Adversarial Network (GAN)

Our Generative Adversarial Network architecture consists of a pre-processing pipeline, a generator model and a discriminator model. (Refer to Figure 2 for a schematic of our GAN architecture.)

To process conditional information, we utilize a pretrained ResNet-18 model to generate image embeddings of identity frames and a pretrained audio encoder to generate feature vectors of audio segments corresponding to each frame. These embeddings are concatenated to serve as the conditional information for the generator. A noise vector also serves as the input of the generator network.

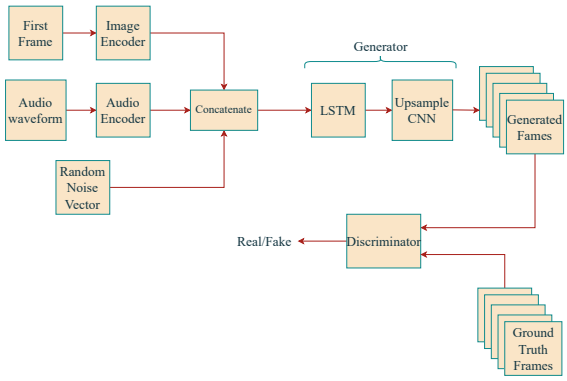


Figure 2: Visual schematic for our conditional GAN architecture.

4.1.1 Generator Network

Our generator is responsible for synthesizing the sequence of video frames. It contains two major components: an LSTM to generate feature vectors for each frame to synthesize, and a small upsampling CNN to transform LSTM features to images.

4.1.2 Discriminator Network

Our discriminator consists of a 3D CNN (with 3D convolution layers and a final linear prediction head) that serves as a video classifier. Its job is to determine whether an input video is real or fake (ie. being generated by the generator.)

4.1.3 GAN Losses

According to [6] and [8], we use the following loss functions for the discriminator $\mathcal{L}_D$ and the generator $\mathcal{L}_G$: [1]

$$\begin{aligned}\mathcal{L}_D = &- \mathbb{E}_{(\mathbf{x}_v,\mathbf{c})\sim p_{\text{data}}}[\log D(\mathbf{x}_v|\mathbf{c})]- \\ &\mathbb{E}_{\mathbf{z}\sim p_z,\mathbf{c}\sim p_{\text{data}}}[\log(1 - D(G(\mathbf{z}|\mathbf{c})|\mathbf{c}))] \\ \mathcal{L}_G = &- \mathbb{E}_{\mathbf{z}\sim p_z,\mathbf{c}\sim p_{\text{data}}}[\log D(G(\mathbf{z}|\mathbf{c})|\mathbf{c})]+ \\ &\lambda_{\text{pixel}}\mathcal{L}_{\text{L1}}(G(\mathbf{z}|\mathbf{c}),\mathbf{x}_v)\end{aligned}$$

In addition to the adversarial term (a qualitative measurement on whether or not the generator was able to fool the discriminator), we also introduce a L1 pixel loss for the generator network $\mathcal{L}_{\text{L1}}$ in order to promote generated frames that are similar to the ground truth:[2]

$$\mathcal{L}\text{L1}(G(\mathbf{z}|\mathbf{c}),\mathbf{x}_v) = \frac{1}{TCHW}\sum_{t=1}^{T}\sum_{c'=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W}|G(\mathbf{z}|\mathbf{c})_{t,c',h,w} - (\mathbf{x}_v)_{t,c',h,w}|$$

4.2. Denoising Diffusion Probabilistic Models (DDPM)

4.2.1 Mathematical Overview

The forward diffusion process gradually adds Gaussian noise to the input image, and is defined in terms of a Markov chain with a pre-defined transition distribution:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) \sim N(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t I)$$

Additionally, an equivalent n-step transition distribution for the forward process can be defined in terms of $\bar{\alpha}_t = \prod_{k=1}^{t}(1-\beta_t)$:

$$q(\mathbf{x}_t|\mathbf{x}_0) \sim N(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)I)$$

The reverse diffusion process is another Markov chain with **learned** Gaussian transition distributions derived from a trained denoising neural network, starting with $p(\mathbf{x}_t) = N(\mathbf{x}_t; 0, I)$: [7]

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \sim N(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t))$$

Our denoising UNet is implemented with the objective of predicting the **noise** being added to the image and subsequently trained using a simplified MSE objective function as a proxy to the variational bound: [7] [3]

$$L_{\text{mse}}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}[||\boldsymbol{\epsilon} - f_\theta(\mathbf{x}_t, t)||^2]$$

4.2.2 Conditional DDPM for Talking Head Generation

Talking head synthesis is an example of a **conditional generation task**. We built our **conditional 2D UNet** based on the unconditional architecture in Dhariwal and Nichol. [5]

Motivated by [14] and [12], our UNet implementation uses two methods of processing conditional information. Conditional information represented by images (identity

[1] x_v is the real sequence of the subsequent frames, c is the conditioning input, z is the noise vector, λ_{pixel} is a hyperparameter that represents the contribution of the pixel loss.

[2] Where H is the height, W is the width, C is the number of channels of a single video frame, and T is the number of video frames.

[3] The original authors used the notation:
$L_{\text{mse}}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)||^2]$,
We substituted the equivalent expression for $\mathbf{x}_t$ for readability and utilize the notation f_θ to emphasize the output of a neural network.

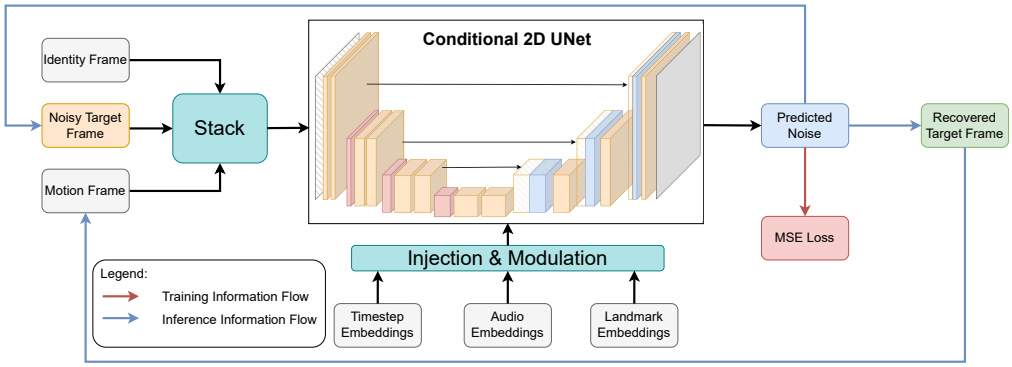


Figure 3: Visual Schematics for Conditional 2D UNet

and motion frames) is fused with the noisy input by stacking along the channel dimension, while conditional information represented by embedding vectors (audio and landmarks) are injected into the UNet's residual blocks alongside with timestep embeddings. Concretely, given the hidden units $\mathbf{x}$ before the "injection", we use an MLP to learn the modulation parameters for embedding z, and we produce

$$\mathbf{x}' = z_s \cdot \text{Norm}(\mathbf{x}) + z_b, \quad \text{where } (z_s, z_b) = \text{MLP}(z)$$

as the modulated hidden units. A visual representation of our UNet architecture is presented in Figure 3.

In addition to the simple MSE loss, we leveraged the extracted facial landmarks to penalize the model's deviation from ground truth noise in the mouth region to promote better lip movement emulation. The penalty is implemented as a weighted MSE loss: The mouth region is weighted by 1.2 while the rest of the image is weighted by 1.0. [14].

4.2.3 Model Configuration and Hyperparameters

We decided to directly utilize the UNet architectures proposed by Dhariwal and Nichol. [5] Our implementations use SiLU activation and GroupNorm normalization, see table 1.

Compared to Dhariwal and Nichol's original architecture, we made the following adjustments in accordance with [12] and [14]:

1. Cosine noise scheduler with 250 diffusion steps during training for both resolutions.
2. Batch size adjusted to fit into GPU memory.
3. Learning rate set to a constant 5e-5 for both resolutions.
4. Spatial self-attention modules only used in the middle block and not in up/down sampling blocks.

In terms of hardware, initial model development was performed with an AMD Radeon Pro 5600M GPU (8 GB), while model training was performed on a GCP node equipped with an Nvidia L4 GPU (24 GB). Both of our 64 x 64 and 128 x 128 models were trained for 200 Epochs. (Totaling 5 GPU days) The 128 x 128 model also incorporates landmark conditioning as a part of our experimentation.

4.2.4 Sampling and Inference

Given that the generation process for a DDPM is formulated as an iterative denoising process, generating a single frame would take a significant amount of time. Two widely used strategies for speeding of the inference process of DDPMs include diffusion timestep spacing and the Denoising Diffusion Implicit Models (DDIM) inference pipeline. [13] We incorporated both in our implementation process, and we mostly used 50 diffusion steps in inference.

Resolution	64	128
Base Channels	192	256
Ch. Multiple	[1,2,3,4]	[1,1,2,3,4]
# Res. Blocks	3	2
Attn. Heads	8	4
Conv. up/down	True	True
Dropout	0.1	0.0
# Parameters	294 Million	470 Million

Table 1: UNet Architectures for DDPM

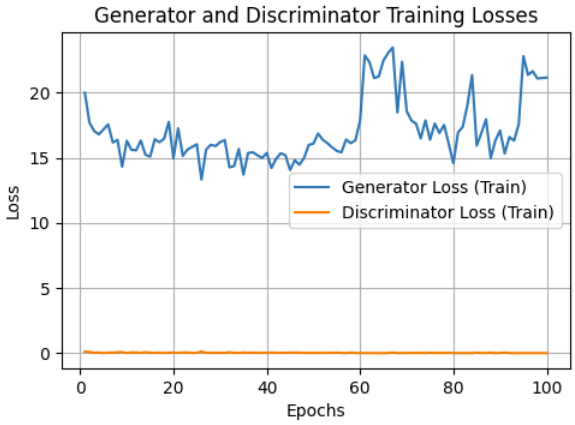


Figure 4: Loss Evolution of the Discriminator and Generator Models During Training

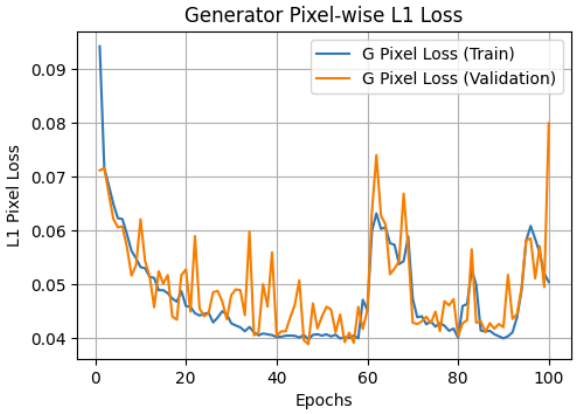


Figure 5: Loss Evolution of the Generator Pixel Loss

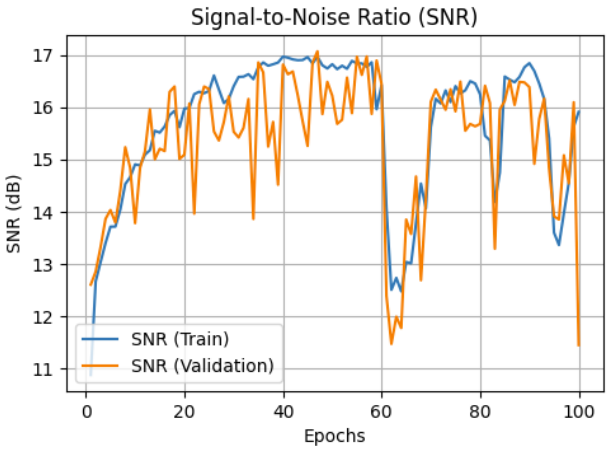


Figure 6: Train and Validation SNRs

Our model generates each frame sequentially and uses the previously generated frame as the motion frame to generate subsequent frames. One major downside of this design choice is that the quality of generated frames would gradually degrade.

5. Experimentation, Results & Discussion

5.1. GAN: Experiments

We will begin by presenting the plots of the training and validation losses for both the generator and discriminator. Although these plots provide a general sense of the model's behavior, they are not very informative on their own. To address this, we added a third loss plot, the pixel loss plot, which is the second part of the generator loss that represents the L1 distance between the generated pixels and the ground truth pixels, offering a more meaningful measure of image quality.

Referring to Figure 4, we see that our discriminator dominated the generator throughout the entire training process. Moreover, the loss with respect to the generator decreased until epoch 60 and started to increase afterwards, which means that the network was suffering from a non-convergence issue as the discriminator completely dominates the generator. Pixel loss presented in Figure 5 is also consistent with this phenomenon, as it started to diverge after epoch 60. In addition, we can see that the validation pixel is much noisier than the training loss, and this is because we don't have many samples in our validation set.

Another useful metric we used to evaluate GANs at training time is the Signal-to-Noise Ratio (SNR) between generated frames and ground truths. Again, Figure 6 is consistent with our existing observations: SNR improved until epoch 60, and then declines due to the generator's outputs becoming noisier and less accurate.

In terms of qualitative observations, our GAN models can only produce videos with mouth movements. The failure to emulate head motions motivated us to experiment with DDPMs in our talking head generation task.

5.2. DDPM: Experiments

5.2.1 Initial Modeling and Experiment Proposals

We present the loss evolution for our base 64^2 diffusion UNet in Figure 7. Two loss characteristics stood out to us. First, the lack of significant gap between training and vali-

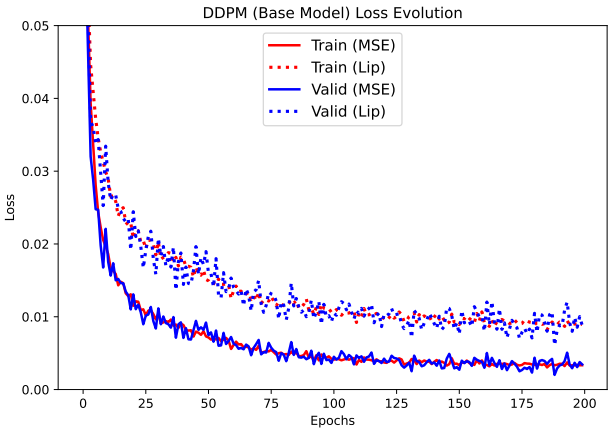


Figure 7: Loss Evolution of Base Diffusion Model (Resolution : 64)

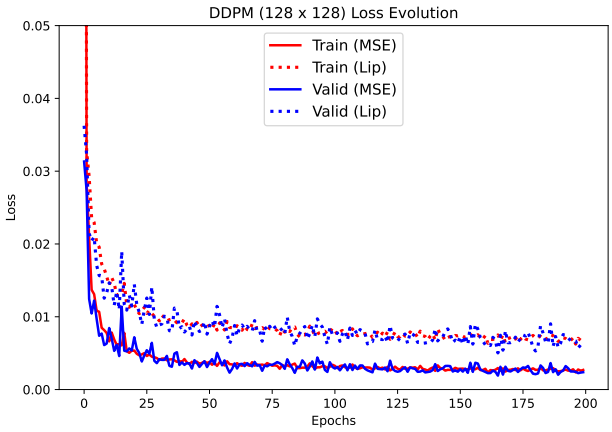


Figure 8: Loss Evolution of 128^2 Diffusion Model with Landmark Conditioning

	0 Epoch (Base)	50 Epoch	100 Epoch
MSE	3.450	2.970	2.310
Lip	8.816	6.940	5.826

Table 2: Validation Losses During Model Finetuning ($\times 10^{-3}$)

dation losses during the entire training process indicates the model's similar denoising performance on both training and validation images. Second, there is a significant difference between unweighted MSE loss and Lip loss, meaning that the model is struggling to capture fine-grained facial details of the actors. Lip loss also converged at a slower rate compared to unweighted MSE.

We proposed two experiments aimed at improving the model's ability to capture finer facial details. First, we increased our image resolution to 128^2 so that the amount of information loss due to image downsampling is reduced. Second, we increased the amount of supervision that encouraged the model's adherence to facial feature consistency by injecting landmark information during training. We present our findings below.

5.2.2 Increasing Image Resolution via Finetuning

Partially motivated by the training approach utilized by the Stable Diffusion model [11], we decided to directly finetune our 64^2 base model on 128^2 images. This fine-tuning experiment lasted for 100 epochs, and some checkpointed validation losses are presented in Table 2.

Even though the loss characteristic suggests that our fine-tuning diminished the gap between Lip loss and unweighted MSE, our qualitative analysis showed that the quality of generated frames significantly decreased. We hypothesize that the model's failure to generalize to higher resolutions may be due to the following:

1. The massive differences in 64^2 and 128^2 models architectures used by [5] suggest that our base model may be ill-suited to process image features in the 128^2 space. (eg. too few model channels)

2. Generating talking faces relies heavily on specific image features that are spatially aligned. Fine-tuning an existing model on higher resolution likely broke the spatial alignment of the features that the model has already established. The upsampling network may become confused about where to put the mouth, eyes etc.

5.2.3 Adding Landmark Conditioning to 128^2 Models

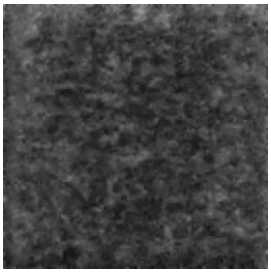
As an effort to improve the generation quality for 128^2, we again adapted the UNet model architecture from [5] and made modifications according to [14] and [12]. This model is also trained with added landmark conditioning. Compared to our base model, the gap between the Lip loss and unweighted MSE is significantly reduced for the 128^2 model. 8

On the other hand, according to generated samples presented in Figure 9, the model is able to generate about 5 frames with consistent facial features and lip movements. However, further frames shows stopped lip movement and gradually increasing brightness.

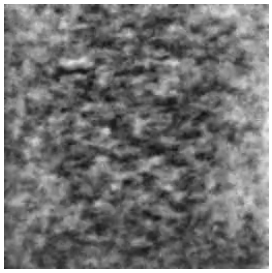
5th Generated Frame

20th Generated Frame

Figure 9: Generated samples from 128^2 model with landmark conditioning



Audio Only



Landmark Only

Figure 10: Relative Strength of Conditioning Signals

Two design choices we made in our modeling methods could potentially contribute to this model's strange behaviors: First, the number of normalization layers in the UNet's residual blocks may be insufficient, [4] which may have caused the UNet to slightly amplify the pixel magnitudes in its predicted noise. This amplification behavior may be too minuscule to be captured by the MSE losses, but it is stacked though the iterative denoising process and creates a feedback loop in the model's inference pipeline. Second, the two control signals being injected into the model do not share the same data modality and thus the underlying embedding distribution, which prompts the phenomenon of control conflicts in the facial region.

To determine the possibility to control conflicts, we generated two test frames with only one control signal given to the model and used the relative pixel magnitude of output frames to determine the strength of each control signal. We saw that lankmarks completely dominated audio embeddings in controlling relative head position and facial feature alignments. 10

[4]There are 2 normalization layers within each residual block, but three feature injection operations are performed.

6. Comparing Results

For subsequent studies and evaluations, we elected our base model (64^2 resolution) checkpoint at 175 epoch as our candidate model based on the criterion of validation loss minimization.

To evaluate the video quality using quantitative metrics, we use Fréchet Inception Distance (FID), which passes the real and generated frames into a pretrained Inception-v3 model and captures the distance between the corresponding feature distributions of the frames. Realistic motion of the generated videos is evaluated using average Optical Flow Magnitude (OFM), which takes the average magnitude of the displacement of pixels between consecutive frames using one channel (grayscale), as required by Farneback, and uses parameters for OpenCV's calcOpticalFlowFarneback as defined in the DiffusedHeads paper [14]. Similarly, to measure the smoothness between frames in a given video, the Frame-wise Mean Square Error (F-MSE) is used, which averages the mse loss across pixels in consecutive frames acrosss all three color channels. We use AV Offset, which is pretrained on the audio-to-video synchronisation network SyncNet [4] to quantitatively measure the difference in time between lip movement and audio. For reference, an AV Offset score of 0 indicates that the lip movement and audio are perfectly synced, while a negative score and positive score, respectively, mean the generated video is lagging or ahead of the audio. AV Confidence indicates the confidence level of a corresponding AV Offset score from the SyncNet model.

6.1. GAN vs Diffusion Model

Although we could not compete with the DiffusedHeads baseline due to computation and cost constraints, we still saw that our diffusion model, although trained on lower resolution (64x64), outperformed the GAN model (128x128) in both video quality (FID, OFM, F-MSE) and audio-video synchronization (AV Offset, Confidence). The GAN has a large AV offset of -3.48 and a large confidence level which indicates the GAN model is confidently wrong to generate the lagging videos, on average, behind the audio. However, because GAN produced blurry frames, it may have confused the SyncNet model's AV offset calculation. The blurriness may have also impacted OFM computation as the GAN model produced a similar OFM score to that of the DiffusedHeads model, even though qualitatively our diffusion model outperformed the GAN. In addition, the evalua-

tion of the OFM and F-MSE score is dependent on the frame resolution. Thus, the OFM and F-MSE scores on 64x64 resolution cannot be compared to that of 128x128. The evaluation of these scores are shown separately in Table**??** and divided by resolution.

6.1.1 Effect of Diffusion Steps at Inference

To investigate the effects of reverse diffusion steps on generation quality, we compared videos generated with 20, 50, and 100 steps using the DDIM pipeline qualitatively and quantitatively. In terms of visual quality, we saw that 50 diffusion steps produced the most realistic and consistent videos, while 20 diffusion steps may sometimes produce talking faces with ill-defined and misaligned facial features. Interestingly, increasing the diffusion steps from 50 to 100 actually reduced the visual fidelity of generated frames. The quantitative results in Table 5 are consistent with our qualitative inspections of Figure 13, specifically that the best video quality, as is captured by FID, is produced at 50 diffusion steps. F-MSE scores across steps were relatively similar across diffusion steps, while FID and AV confidence had more variability. Our diffusion model's small AV Offset of 0.16 at 50 diffusion steps demonstrated it was the most audio synced option and furthermore had a highest confidence score in this quantitative assessment. (More in-depth Comparison samples are included in the appendix.)

7. Conclusion and Future Works

In this project, we experimented Generative Adversarial Networks (GANs) and Diffusion Models (DDPMs) on the generative computer vision task of talking head synthesis. We implemented our bespoke GAN network architecture and adapted an existing denoising UNet architecture to our task by combining two methods of conditional signal processing. In our analysis, we also showed that using 50-step DDIM sampling method during inference achieves the best balance between computational efficiency and sample quality. Our best performing model is not only able to produce talking head animations with accurate lip sync but also can emulate head movements of the talking subject.

Despite the overall results of our models, our findings call for further work. For instance, we wish to increase the capacity of the generator network in our GAN so that the non-convergence issue and discriminator dominance could be addressed, and we could implement conditional signal combination strategies that avoid control conflict, such as Diffusion Mamba (DiM). [15] Moreover, harnessing sophisticated loss metrics, such as such as Learned Perceptual Image Patch Similarity (LPIPS) [19] and identity preserving losses, would improve the overall generation quality. Finally, Finally, our dataset only contains 91 different face identities and has a limitation of green backgrounds, which inhibits our models to generalize to generic faces and identities. We believe background data augmentation and more diverse datasets such as MEAD [17] and Lip Reading in the Wild (LRW) [4] could help with generalization performance.

Model	FID ↓	AV Offset	AV Conf. ↑
GAN (128x128)	94.255	-3.48	2.682
Diffusion (64x64)	81.682	0.16	2.436
GT (64x64)	0	0.88	3.306
Baseline: DiffusedHeads	49.03	0.88	2.91

Table 3: Quantitative metrics of GAN vs Diffusion (note: DiffusedHeads metrics were normalized to our scale by the ratio of the respective ground truth of the resolution used)

Model/Resolution	OFM	F-MSE ↓
GAN (128x128)	0.680	31.688
Baseline: DiffusedHeads (128x128)	0.643	6.99
GT (128x128)	0.690	7.76
Diffusion (50 steps, 64x64)	0.342	37.014
GT(64x64)	0.534	54.433

Table 4: OFM and F-MSE Analysis (relative to 64x64 and 128x128 resolution comparison)

# Steps	FID ↓	OFM	F-MSE ↓	AV Offset	AV Conf. ↑
20	84.968	0.338	37.887	-0.28	1.895
50	81.682	0.342	37.014	0.16	2.436
100	88.366	0.343	37.872	-0.32	2.061
GT	0	0.534	54.433	0.88	3.306

Table 5: Quantitative metrics diffusion models with different number of DDIM inference steps on 64x64.

References

[1] BARNETT, S. A. Convergence problems with generative adversarial networks (gans), 2018.

[2] BRAND, M. Voice puppetry. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques* (USA, 1999), SIGGRAPH '99, ACM Press/Addison-Wesley Publishing Co., p. 21–28.

[3] CAO, H., COOPER, D. G., KEUTMANN, M. K., GUR, R. C., NENKOVA, A., AND VERMA, R. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing 5*, 4 (2014), 377–390.

[4] CHUNG, J. S., SENIOR, A., VINYALS, O., AND ZISSERMAN, A. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017), IEEE.

[5] DHARIWAL, P., AND NICHOL, A. Diffusion models beat gans on image synthesis, 2021.

[6] GOODFELLOW, I. J., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A., AND BENGIO, Y. Generative adversarial networks, 2014.

[7] HO, J., JAIN, A., AND ABBEEL, P. Denoising diffusion probabilistic models, 2020.

[8] ISOLA, P., ZHU, J.-Y., ZHOU, T., AND EFROS, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).

[9] PENG, Y. A comparative analysis between gan and diffusion models in image generation. *Transactions on Computer Science and Intelligent Systems Research 5* (08 2024), 189–195.

[10] PRAJWAL, K. R., MUKHOPADHYAY, R., NAMBOODIRI, V. P., AND JAWAHAR, C. 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* (Oct. 2020), MM '20, ACM, p. 484–492.

[11] ROMBACH, R., BLATTMANN, A., LORENZ, D., ESSER, P., AND OMMER, B. High-resolution image synthesis with latent diffusion models, 2022.

[12] SHEN, S., ZHAO, W., MENG, Z., LI, W., ZHU, Z., ZHOU, J., AND LU, J. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation, 2023.

[13] SONG, J., MENG, C., AND ERMON, S. Denoising diffusion implicit models, 2022.

[14] STYPUŁKOWSKI, M., VOUGIOUKAS, K., HE, S., ZIEBA, M., PETRIDIS, S., AND PANTIC, M. Diffused heads: Diffusion models beat gans on talking-face generation, 2023.

[15] TENG, Y., WU, Y., SHI, H., NING, X., DAI, G., WANG, Y., LI, Z., AND LIU, X. Dim: Diffusion mamba for efficient high-resolution image synthesis, 2024.

[16] VOUGIOUKAS, K., PETRIDIS, S., AND PANTIC, M. End-to-end speech-driven facial animation with temporal gans, 2018.

[17] WANG, K., WU, Q., SONG, L., YANG, Z., WU, W., QIAN, C., HE, R., QIAO, Y., AND LOY, C. C. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV* (2020).

[18] XIE, L., AND LIU, Z.-Q. A coupled hmm approach to video-realistic speech animation. *Pattern Recognition 40*, 8 (2007), 2325–2340. Part Special Issue on Visual Information Processing.

[19] ZHANG, R., ISOLA, P., EFROS, A. A., SHECHTMAN, E., AND WANG, O. The unreasonable effectiveness of deep features as a perceptual metric, 2018.

[20] ZHOU, Y., HAN, X., SHECHTMAN, E., ECHEVARRIA, J., KALOGERAKIS, E., AND LI, D. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics 39*, 6 (Nov. 2020), 1–15.

Contributions

In general, all team members worked together in the stages of literature review, dataset selection, model architecture study, experimentation, and result collection. Additionally, we collaborated to streamline the ML infrastruc-

ture and cross-checked for errors. For contribution for specific tasks, consult the list below.

1. Literature review: Songyu, Mustafa, Hannah (Equally)
2. Dataset sourcing: Mustafa and Hannah
3. Dataset preprocessing: Songyu
4. GAN model development and training: Mustafa
5. DDPM model and DDIM inference pipeline development: Songyu
6. DDPM model training and experimentation: Songyu and Hannah
7. Evaluation metrics development and experimental data analysis: Hannah
8. Report: Songyu, Mustafa, and Hannah (Equally)

8. Appendix

Exhibition of Generated Samples

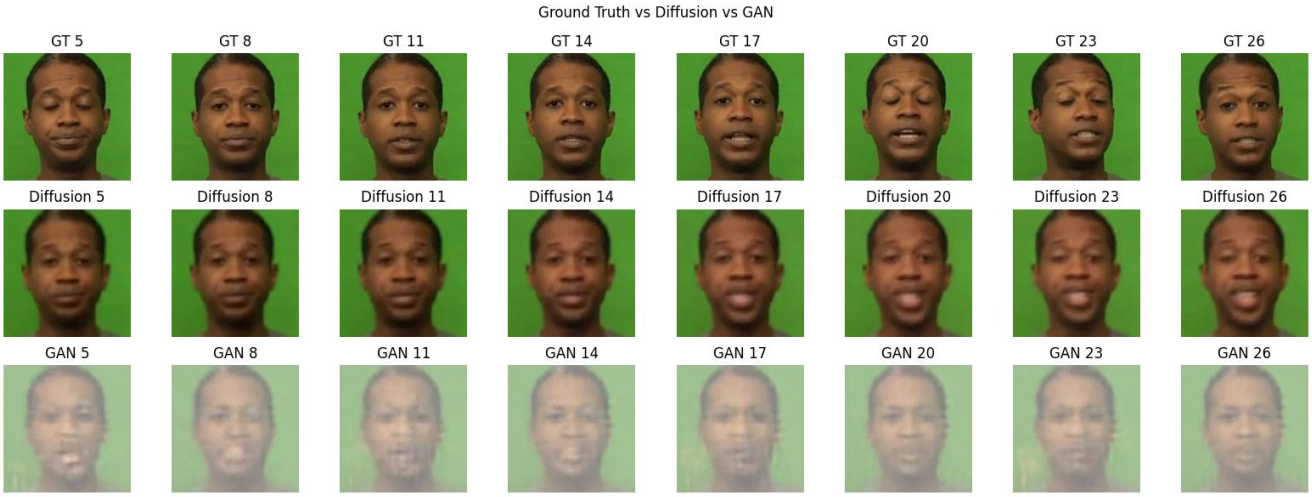


Figure 11: Comparison Between our GAN candidate and Diffusion candidate, Sample 1

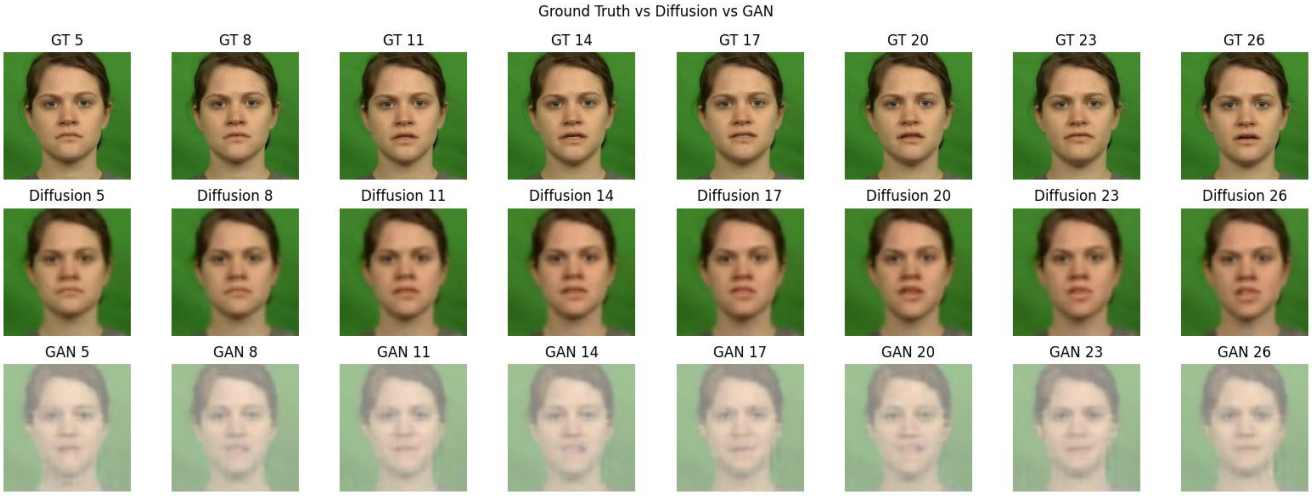


Figure 12: Comparison Between our GAN candidate and Diffusion candidate, Sample 2

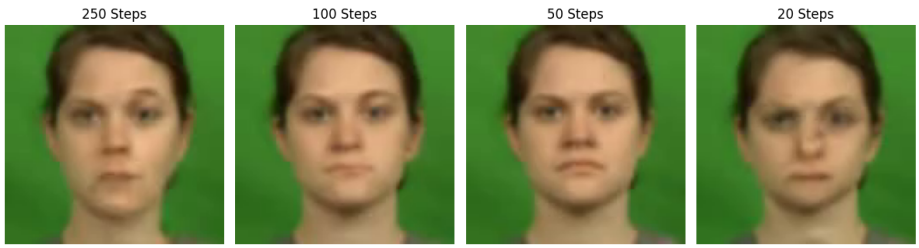


Figure 13: Qualitative Analysis of Diffusion Steps for First Frame