Laughing Matters: Introducing Laughing-Face Generation using Diffusion Models

Antoni Bigata Casademunt
a.bigata-casademunt22@imperial.ac.uk
Rodrigo Mira
rs2517@imperial.ac.uk
Nikita Drobyshev
nikita.drobyshev23@gmail.com
Konstantinos Vougioukas
k.vougioukas@imperial.ac.uk
Stavros Petridis
stavros.petridis04@imperial.ac.uk
Maja Pantic
m.pantic@imperial.ac.uk

Intelligent Behaviour Understanding Group (iBUG)
Imperial College London, UK

Abstract

Speech-driven animation has gained significant traction in recent years, with current methods achieving near-photorealistic results. However, the field remains underexplored regarding non-verbal communication despite evidence demonstrating its importance in human interaction. In particular, generating laughter sequences presents a unique challenge due to the intricacy and nuances of this behaviour. This paper aims to bridge this gap by proposing a novel model capable of generating realistic laughter sequences, given a still portrait and an audio clip containing laughter. We highlight the failure cases of traditional facial animation methods and leverage recent advances in diffusion models to produce convincing laughter videos. We train our model on a diverse set of laughter datasets and introduce an evaluation metric specifically designed for laughter. When compared with previous speech-driven approaches, our model achieves state-of-the-art performance across all metrics, even when these are re-trained for laughter generation. Our code and project are publicly available [1].

1 Introduction

Facial animation is essential in many applications, such as virtual reality, movies, and human-computer interaction (HCI), by providing more immersive and engaging experiences. Current facial animation methods predominantly focus on speech-driven animation, resulting in

[1] https://sites.google.com/view/laughing-matters

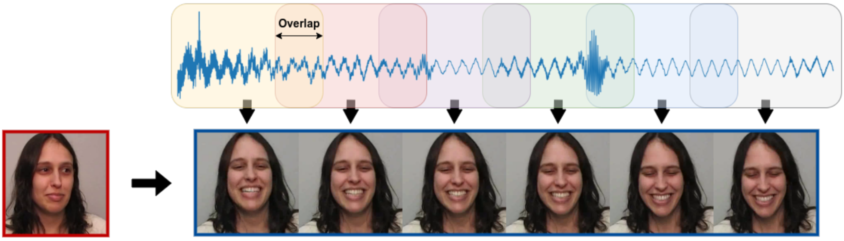

Figure 1: The proposed end-to-end laughter generation model synthesizes a video of a laughing face using a still image of the speaker and an accompanying laughter segment.

the neglect of non-verbal expressions such as laughter, head nods, or blinks. This omission poses a substantial issue since these non-verbal cues often convey essential contextual information and play an important role in natural dialogue. Laughter is an interesting initial subject of study due to its ancient roots as a social signal [35, 36, 41], acting as a powerful non-verbal communication medium that conveys emotions, intentions, and social relationships [15, 39]. However, laughter, unlike speech, lacks a direct correlation with lip movement. When combined with the scarcity of training data, this makes the development of a model for realistic laughter sequence generation quite challenging.

Until recently, speech-driven animation methods mainly relied on Generative Adversarial Networks (GANs) [16]. Early approaches were limited in terms of head rotations [55] or could only modify lip movements [38]. Recent advances have led to methods capable of generating realistic facial animations, with [32, 61, 62] or without [56] the use of intermediate representations such as key points, landmarks, or driving videos. Some of these methods even incorporate emotion control into the generation process [17, 23]. The emergence of diffusion-based generation techniques has further spurred progress in the field, as researchers leverage the improved performance of these new models [44, 50]. Current methods employ frame-based generators, exploiting the strong correlation between speech and lip movement. However, these models struggle with laughter generation due to several issues. Firstly, laughter lacks the robust audio-visual correlation seen in speech [1, 24], making the generation of authentic audio-driven laughter sequences considerably more difficult. Secondly, laughter's complexity and variability, involving various muscles and facial movements, poses a substantial challenge for existing frame-based generators. These models, which are designed for speech, primarily focus on the mouth and lips and struggle to capture the subtleties and variations in laughter, resulting in unnatural or inaccurate visual renditions. Finally, the spontaneity and context-dependency of laughter make it difficult to predict the timing and intensity of the speaker's facial movements accurately. These challenges emphasize the need for innovative approaches specifically designed for laughter generation.

In this paper, we design a novel video diffusion model to generate videos of laughing faces based on raw audio input. Our model leverages recent developments in video diffusion [21, 45] to accurately capture the complex laughter dynamics, leading to realistic and synchronized laughing animations. To the best of our knowledge, our method is the first to generate natural laughter videos. In addition, we address the issue of limited publicly-available audio-visual laughter corpora by proposing an ensemble of existing datasets for training and evaluation purposes. To assess the quality of our results, we employ a series of metrics from existing video generation works and design a novel metric specifically tailored for laughter generation. We perform a thorough evaluation of our proposed method's performance and conduct an ablation study on our model to systematically assess the impact of each individual component within this system. We find that our approach outperforms previous state-of-the-art speech-driven facial animation models, including other diffusion-based methods, whether pre-trained on speech or re-trained on laughter. Furthermore, our method produces videos that are significantly better aligned with the input laughter audio.

2 Related Work

Speech-Driven Facial Animation. Early facial animation research [60] established a strong relationship between speech features and facial motion, initially leveraged through hidden Markov models (HMMs) [58, 59]. Quality enhancements were realized with the rise of deep learning methods [13, 25, 51], notably through the introduction of generative adversarial networks (GANs) [16]. As a result, an initial wave of research focused on achieving lip synchronization [6, 51] with Prajwal *et al.* [38] attaining near-perfect synchronization for in-the-wild videos. Subsequent work incorporated natural facial expressions such as blinks and eyebrow movements but lacked head rotations [46, 56]. Some methods addressed this by using intermediate representations like landmarks [17, 62], keypoints [23] or a driving video [32, 61]. On the other hand, recent diffusion-based approaches [50] have demonstrated state-of-the-art performance, showcasing their ability to generate plausible head motion and diverse facial expressions by using only speech as a conditioning input. Other works have also focused on adding control over the emotion displayed in the generation [1, 17, 32]. Despite these remarkable advances, the generation of non-verbal aspects of human communication remains unexplored.

Diffusion Models. Introduced in [19, 47, 48], diffusion models have shown strong generative capabilities in point cloud generation [34], music synthesis [22, 42] and video generation [20, 21, 45]. Compared to GANs, diffusion models provide a more stable and robust training process, as well as improved mode coverage which makes the model less likely to overfit [57]. Score-based diffusion models, presented in [49] and improved in [8, 27], extend the original diffusion models by generalizing the noise distribution through the use of stochastic differential equations (SDEs). They can effectively capture complex data distributions and generate high-quality samples, while still maintaining the advantages of the original diffusion models in terms of denoising and sampling efficiency. These advancements allow for a broader range of applications and adaptability to different domains. More recently, [40] propose Latent Diffusion Models, managing to produce high-resolution images by transferring the training and inference processes to a compressed lower-dimension latent space for more efficient computing. However, we found that this approach failed to yield successful results in our case, likely due to the limited amount of data available.

Laughter in Human Communication. Laughter generation has been explored across various modalities, including audio, text, and animation. In the audio domain, studies have primarily focused on synthesizing laughter sounds [30, 52, 54] to extend the capabilities of text-to-speech systems. In the text domain, researchers have investigated methods for generating and recognizing laughter in textual conversations, such as identifying and generating laughter events in dialogues [2]. For animation, Ding *et al.* [9] developed a real-time laughter animation generator that takes input pseudo-phonemes of laughter and their duration times, synthesizes facial and body motions by learning the relationship between input signals and human motions, and employs a combination of contextual Gaussian Models and motion

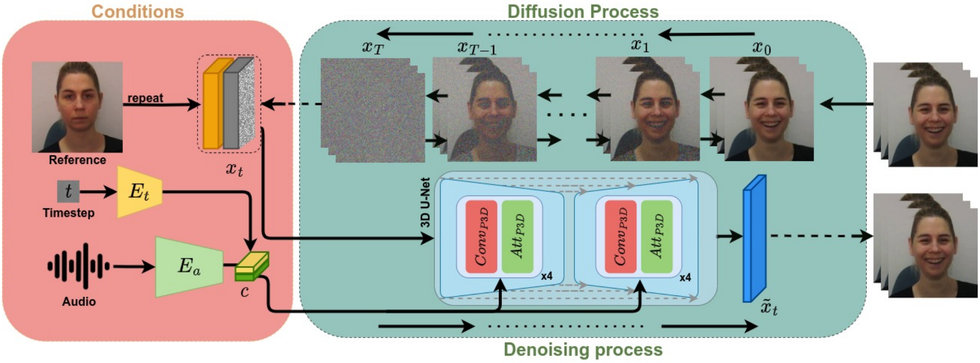

Figure 2: Overview of our proposed pipeline for laughter generation. The model takes as input the noisy video concatenated with the reference frame and outputs the denoised version of the images conditioned on the laughter audio and the timestep of the diffusion process.

capture data. More recently, projects like ILHAIRE [12] have showcased the importance of laughter synthesis and recognition in human-avatar interactions. Incorporating laughter in facial animation is crucial for developing more realistic and engaging virtual characters, ultimately enhancing the overall user experience.

3 Methodology

3.1 Diffusion models

Diffusion models [19, 47, 48] are a class of generative models that synthesize samples by progressively removing random noise. The input to a conditional diffusion model consists of a conditioning signal c, a random time step t, and a sample x_t obtained by corrupting the original data x by adding i.i.d. Gaussian noise of standard deviation σ.

We adopt the approach of Karras *et al.* [27] that further explores the design choices of this type of model, both theoretically and empirically, and presents a sampling process that uses Heun's method as the ODE solver, reducing the number of neural function evaluations needed while maintaining the FID score [7]. This process is characterized by a noise schedule with a standard deviation σ_t at time t. The time range t is uniformly sampled during training, with the diffusion progressing in the direction of increasing time. The Gaussian diffusion dynamics can be fully described by a single noise vector $\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ with noise levels $\sigma_0 = \sigma_{max} > \sigma_1 > \cdots > \sigma_T = 0$, as x_t can be expressed as a function of the original sample and the noise vector $\boldsymbol{n}$, i.e., $x_t = x + \boldsymbol{n}$. The model D_θ is trained to determine the original image given this input. The diffusion loss minimizes the expected L_2 denoising error for samples drawn from the training data separately for every σ, i.e.:

$$\mathcal{L}(\theta) = \mathbb{E}_{x,c,t,\sigma}[w_t \| D_\theta(x_t; c, \sigma_t) - x\|_2^2], \tag{1}$$

where w_t is a fixed weight function of choice. Inference is performed by taking random noise at time t_{max} and denoising it using the noise predictions provided by the model.

3.2 Architecture

Our architecture primarily builds upon the work of Ho *et al.* [21], which employs a factorized space-time U-Net architecture, extending the standard 2D U-Net used in image diffusion models. The model, illustrated in Fig. 2, comprises four down-sampling and up-sampling blocks connected by residual connections. The input of our model is a video sample $x_t \in \mathbb{R}^{B\times C\times F\times H\times W}$, where B, C, F, H, W are the batch, channels, frames, height, and width dimensions respectively. The condition signal c, in our case, consists of a single frame $x' \in \mathbb{R}^{B\times C\times H\times W}$ concatenated channel-wise with x_t by repeating it in the temporal dimension, and an audio sequence a which is passed through an audio encoder (E_a). Additionally, we pass the timestep information of the diffusion process t processed by a two-layer MLP (E_t). The U-Net contains a composition of convolutional and self-attention layers followed by a down-sampling or up-sampling layer. In our proposed method, we apply Pseudo-3D Convolutional and Attention Layers [45] to balance computational efficiency, and information sharing in the network. For each layer, instead of using the full 3D convolution, we use a 2D convolution applied to the spatial dimensions ($\mathbb{R}^{B\times C\times F\times H\times W} \rightarrow \mathbb{R}^{(B\times F)\times C\times H\times W}$), followed by a 1D convolution applied to the temporal dimension by merging the other dimensions ($\mathbb{R}^{B\times C\times F\times H\times W} \rightarrow \mathbb{R}^{(B\times H\times W)\times C\times F}$). We apply a similar strategy for the attention layers. The details of the network can be found in Fig. 3.

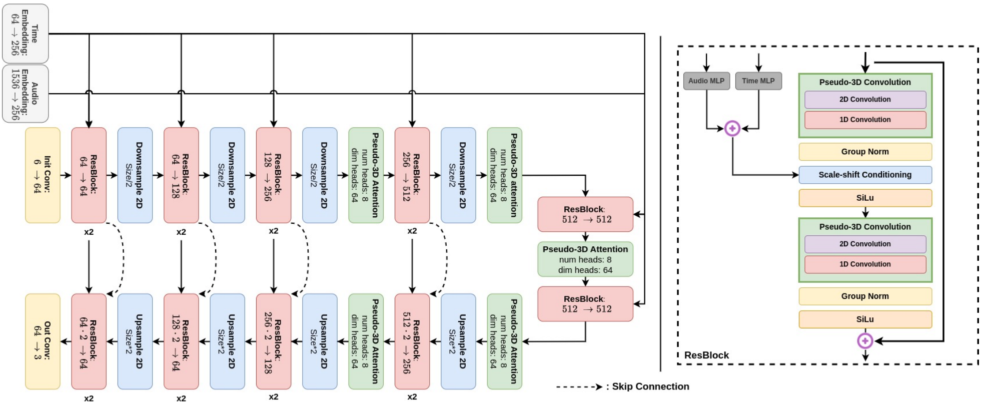

Figure 3: Details of the U-Net layers. Left: Overview of the U-Net architecture. Right: Detailed view of the ResBlock.

The audio signal is processed using an audio encoder from [4] pre-trained on AudioSet [14]. We split the corresponding audio sequence into chunks of equal length based on the number of frames in the video, resulting in a new audio sequence $a' = \{a'_0, ..., a'_F\}$. Then, the audio and timestep conditioning is performed in each of the ResBlocks of the U-Net at the first convolutional layer. This is done by modulating the input x_t through a scale-shift operation after a group normalization layer (GN):

$$h_{s+1} = GN(h_s) * (k+1) + s \tag{2}$$

where h_s and h_{s+1} are consecutive hidden states of the U-Net, and k and s are the scale and shift, respectively. To obtain k and s, we sum the audio sequence a_t and the encoded timestep information t and pass it through a linear layer. We then split the result to obtain the scale and shift.

3.3 Training

One challenge in generating laughter is the lack of large publicly available datasets. To mitigate the risk of overfitting, a common issue in diffusion models trained on smaller datasets, we adopt a series of techniques described below that we further discuss in section 5.2:

Augmentation regularization. We use a technique originally developed for training GANs with limited data [26] and later successfully applied to diffusion models [27]. The pipeline incorporates several geometric transformations which are applied to training images prior to adding noise. To prevent these augmentations from leaking into the generated images, we supply the augmentation parameters as conditioning input to D_θ. During inference, we set these parameters to zero, ensuring only non-augmented images are generated.

Classifier-free guidance (CFG). This technique, proposed by Ho and Salimans [18], enhances the alignment between generated samples and conditional input. At inference, the noise vector is computed both with and without the conditional input, and the final noise vector is determined by $wD_\theta(x_t; c, \sigma_t) + (1-w)D_\theta(x_t; \cdot, \sigma_t)$, where $w > 1$. We found that $w = 1$ worked best in our case. During training, the conditional input is masked with a probability of 0.1, enabling the same model to handle both conditional and unconditional generation.

Longer sequence generation. Due to computational constraints, we train our model on sequences of 16 consecutive frames randomly sampled from the videos, rather than training on the full videos. However, during inference, we can generate arbitrary lengths by autoregressively sampling our model. Initially, we generate a video $x_a \sim p_\theta(x)$ and use the last frame of x_a as a condition for $x_b \sim p_\theta(x_b|x_a)$.

We train all models using the Lion optimizer [5] with $\beta_1 = 0.95$ and $\beta_2 = 0.98$ and a learning rate of 6×10^{-5}. During initial experimentation, we found that Lion converged noticeably faster than commonly used optimizers such as Adam [29] or AdamW [33], while consistently achieving equivalent or superior final performance. We linearly warm up the learning rate for the first 20 epochs and subsequently apply a cosine decay schedule until the end of training. We train our models for 200 epochs with a total minibatch of 32 samples.

4 Experiments

4.1 Datasets

We identified four datasets suitable for laughter generation, namely MAHNOB [37], AVLaughterCycle [53], AVIC [43] and SAL [10]. As these datasets do not solely contain laughter, we focus on the videos that feature it. We divide the data into training, validation, and test sets following an 80 – 10 – 10 % ratio, ensuring there is no overlap between the speakers in each set. The exact split of data can be found in the supplementary material. For all datasets, we use an audio sampling rate of 16 kHz and a video frame rate of 25. During preprocessing, we align all faces to a canonical face and normalize images to the [-1,1] range. Details for each dataset are presented in Table 1.

4.2 Evaluation Metrics

We employ widely-used reconstruction metrics, such as peak Frechet Inception Distance (FID) and structural similarity (SSIM) index, to assess the quality of generated images. Furthermore, we employ Frechet Video Distance (FVD) to evaluate visual quality, temporal

Dataset	# Speakers	# Videos	Avg. length (sec.)	Total length (hours)
AVLaughterCycle [53]	8	421	3.80 ± 6.42	0.44
Mahnob [37]	22	554	1.56 ± 2.21	0.24
AVIC [43]	21	312	0.36 ± 0.30	0.03
SAL [10]	28	98	1.46 ± 0.77	0.04

Table 1: Overview of the datasets used in the study.

coherence, and sample diversity. To assess the authenticity of the generated laughing faces, we train a Laughter Classifier (LC) to differentiate between speech and laughter videos. This model, based on a pre-trained MViTv2 [31] for video classification on Kinetics [28], is fine-tuned with laughter and speech data from MAHNOB [37]. More details are available in the supplementary material. This novel metric allows us to highlight the limitations of pre-trained speech-driven animation methods, while simultaneously demonstrating our model's capability to generate realistic laughter sequences. The Laughter Classifier achieves an accuracy and F1 score of 100 % on the test set. We then apply this trained model to categorize the generated videos, assessing whether they are accurately classified as laughter. When measuring the Laughter Classifier's accuracy, we prevent any bias caused by the initial frame by ensuring that the reference frame is a neutral face, which may not always be the case when sampling a random video. This is crucial as a smiling face can easily resemble laughter, introducing bias into our evaluation.

5 Results

To the best of our knowledge, this is the first work on audio-driven laughter generation, so we compare against three speech-driven animation methods that we re-trained for laughter: Diffused Heads [50], SDA [55], and EAMM [23]. We also compare with pre-trained models such as MakeItTalk [62] and PC-AVS [61]. Furthermore, we perform ablation studies on various design choices within our framework and discuss their importance. In terms of video generation, we evaluate the models at a resolution of 128 × 128. These models are conditioned on a single frame and generate the following 16, adhering to the FVD model's expectation of a 16-frame video. However, for human evaluations, we opted to generate 2-second videos.

5.1 Comparison with Other Works

As shown in Table 2, models that are pre-trained on speech struggle to generate satisfactory results, especially in terms of the Laughter Classifier metric. This highlights the need for re-training the models on laughter. Consequently, we also compare our method against re-trained models, adhering to the recommended parameters from their respective papers. Despite the improvements achieved by re-training, our approach consistently outperforms other methods in terms of visual quality and laughter accuracy. We primarily attribute our model's performance to our 3D architecture, which, unlike other frame-based methods, enables longer audio context. This is essential since laughter exhibits a lower correlation between acoustic and visual cues compared to speech. Other significant improvements stem

Model	FVD ↓	FID ↓	SSIM ↑	LC ↑ (%)	MOS ↑
Pre-trained					
Diffused Heads [50]	149.51	49.36	0.236	80.70	-
SDA [55]	594.32	111.89	0.053	13.85	-
EAMM [23]	391.62	71.71	0.094	16.67	-
PC-AVS [61]	1164.49	175.99	0.004	53.91	-
MakeItTalk [62]	196.89	49.08	0.262	72.50	1.94±1.12
Re-trained					
Diffused Heads [50]	152.30	67.46	0.232	94.09	2.45±1.22
SDA [55]	696.33	124.52	0.040	85.13	-
EAMM [23]	324.97	74.18	0.095	20.67	1.87±1.05
Laughing Matters (Ours)	**111.95**	**45.69**	**0.371**	**96.52**	**3.39±1.09**
Ground truth	-	-	-	100.00	3.49±1.23

Table 2: Comparative performance of the proposed methods against pre-trained and re-trained models. The best result is highlighted in **bold**.

from the choice of an audio encoder specifically tailored for laughter and the training improvements employed to compensate for limited training data, as discussed in Section 5.2.

We further validate our model's superior performance through a Mean Opinion Score (MOS) test. Participants are shown an average of 23 randomly selected videos, featuring a blend of ground truth, our model, Diffused Heads [50], MakeItTalk [62], and EAMM [23]. Participants watch the videos sequentially and rate them on a scale of 1 to 5, where 1 indicates the video appears clearly artificial, and 5 suggests it is highly realistic and indistinguishable from genuine laughter. We collect a total of 72 responses that we detail in the supplementary material. Even though the performance difference is minimal in the laughter metric, it is significant in terms of user preference, where temporal smoothness and natural expressions are crucial factors. It is worth highlighting that even ground truth videos score relatively low, which is likely due to the difficulties in assessing whether a laughter video is realistic and well synchronized with its corresponding audio. This discriminative task is indeed challenging, even for humans, as evidenced by the videos provided in the supplementary material.

Additionally, we compare two variants of our approach as two different models in the user study: with and without head rotations. This is achieved by taking the original video and eliminating the head rotation using the model from [11]. We evaluate both models on videos taken from the MAHNOB [37] test set. The results, shown in Table 3, show that removing head rotation severely deteriorates performance, highlighting the importance of correctly modelling the speaker's head movements when generating laughter videos.

Model	MOS ↑
Laughing Matters w/ rotations	3.08 ± 1.12
Laughing Matters w/o rotations	2.08 ± 1.07

Table 3: Mean Opinion Score of our model with and without the head rotations.

5.2 Ablation study

Audio Encoder. Choosing the right audio encoder is essential for achieving optimal results. While a pre-trained model on a large speech dataset is a common choice for speech animation, they prove inadequate for our specific use case. As indicated by Table 4, speech encoders such as SDA [55] and WavLM [3] yield unsatisfactory results, producing outputs closer to speech rather than laughter, as observed in the Laughter Classifier metric. Training from scratch, for instance with mel-spectrograms, provides some improvement as it allows the model to learn directly from the laughter data. However, due to the limited availability of training data, it is highly beneficial to identify a pre-trained model suitable for our task. To this end, we apply BEATs [4], a state-of-the-art self-supervised audio encoder. Being trained on AudioSet [14], which contains 15.8 hours of laughter data, it achieves superior performance across all metrics.

Audio Encoder	FVD ↓	FID ↓	SSIM ↑	LC ↑ (%)
SDA [55]	169.48	55.07	0.318	68.21
WavLM [3]	136.76	46.01	0.312	54.21
Mel-spectrograms	124.81	47.74	0.320	83.52
BEATs [4]	**111.95**	**45.69**	**0.371**	**96.52**

Table 4: Ablation study on the audio encoder.

Training improvements. To mitigate overfitting, a common issue with diffusion models trained on smaller datasets, we implement two techniques detailed in Section 3.3: Augmentation regularization and Classifier-free guidance (CFG). Table 5 illustrates the impact of both components, demonstrating consistent performance improvements when they are used.

Training configuration	FVD ↓	FID ↓	SSIM ↑	LC ↑ (%)
Baseline	**111.95**	**45.69**	**0.371**	**96.52**
w/o Augmentation regularization	195.03	60.60	0.308	83.93
w/o Classifier-free guidance	126.89	46.91	0.302	75.09

Table 5: Ablation study on the training improvements.

5.3 Qualitative evaluation

Our method effectively generates realistic videos from previously unseen faces and audio clips taken from the test set. Fig. 4 illustrates the laughter sequence generated by our model and competing approaches. Upon visual examination, it is apparent that EAMM [23] struggles to preserve identity, whereas MakeItTalk [62] only animates the lips. While Diffused Heads [50] can consistently produce high-quality visuals, the synchronization with the audio input often falls short. Conversely, our model succeeds in creating a laughter sequence with correlated head movement. For a deeper understanding of our results, we invite readers to review the generated videos available in the supplementary material, where our model generates various laughter types and ad- justs to out-of-distribution speakers.

Moreover, our aim is to demonstrate that our model can replicate the movement patterns seen in real laughter videos. Fig. 5 presents a comparison of the average magnitude of

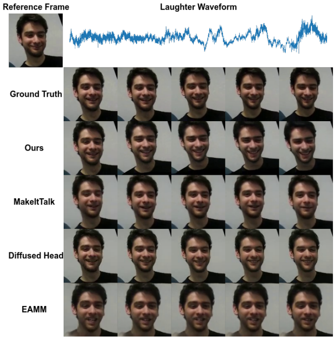

Figure 4: Qualitative evaluation results. The reference frame and the laughter waveform can be seen on the top.

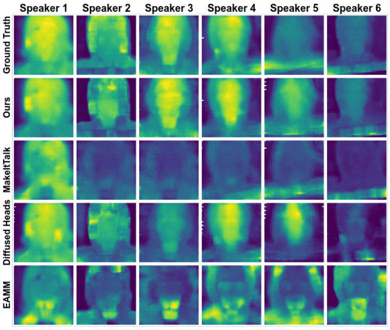

Figure 5: Average magnitude of optical flow for different speakers across all videos.

optical flow for various speakers, showing the regions of the frames that exhibit the most movement. The heatmaps from our generated videos closely align with the ground truth across all speakers, validating our model's ability to create laughter sequences with natural movement. In contrast, MakeItTalk [62] and EAMM [23] yield results that significantly deviate from the ground truth. Notably, while Diffused Heads [50] generates somewhat accurate movements, it falls short in matching the ground truth for Speakers 3, 5, and 6.

6 Conclusion

In this work, we introduce Laughing Matters, an end-to-end model that synthesizes realistic laughing faces from a still image and an audio clip. Our approach outperforms existing methods in generating convincing laughter animations, as demonstrated through quantitative and qualitative evaluations. We conduct a set of ablation studies to examine the impact of the audio encoder and training improvements. Our findings reveal that using a laughter-specific audio encoder, applying augmentation regularization techniques, and leveraging classifier-free guidance significantly enhance the model's performance. Looking forward, it would be promising to extend our model to cover other non-verbal cues, with the aim of creating a comprehensive facial animation model that can animate all verbal and non-verbal cues present in natural speech.

References

[1] Triantafyllos Afouras, Honglie Chen, Weidi Xie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Audio-visual synchronisation in the wild. In *BMVC*, page 261. BMVA Press, 2021.

[2] Dario Bertero and Pascale Fung. A long short-term memory framework for predicting humor in dialogues. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 130–135, 2016.

[3] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16:1–14, 10 2022. doi: 10.1109/JSTSP.2022.3188113.

[4] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. *CoRR*, 12 2022. doi: 10.48550/arXiv.2212.09058.

[5] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms. *CoRR*, abs/2302.06675, 2023.

[6] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? In *BMVC*. BMVA Press, 2017.

[7] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

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

[9] Yu Ding, Ken Prepin, Jing Huang, Catherine Pelachaud, and Thierry Artières. Laughter animation synthesis. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 773–780, 2014.

[10] Ellen Douglas-Cowie, Roddy Cowie, Cate Cox, Noam Amir, and Dirk Heylen. The sensitive artificial listener: an induction technique for generating emotionally coloured conversation. In *LREC workshop on corpora for research on emotion and affect*, pages 1–4. ELRA Marrakech, Morocco, 2008.

[11] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *ACM Multimedia*, pages 2663–2671. ACM, 2022.

[12] Stéphane Dupont, Hüseyin Çakmak, Will Curran, Thierry Dutoit, Jennifer Hofmann, Gary McKeown, Olivier Pietquin, Tracey Platt, Willibald Ruch, and Jérôme Urbain. Laughter research: a review of the ilhaire project. *Toward Robotic Socially Believable Behaving Systems-Volume I: Modeling Emotions*, pages 147–181, 2016.

[13] Bo Fan, Lijuan Wang, Frank K Soong, and Lei Xie. Photo-real talking head with deep bidirectional lstm. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4884–4888. IEEE, 2015.

[14] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[15] Phillip Glenn. *Laughter in interaction*, volume 18. Cambridge University Press, 2003.

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

[17] Siddharth Gururani, Arun Mallya, Ting-Chun Wang, Rafael Valle, and Ming-Yu Liu. Spacex: Speech-driven portrait animation with controllable expression. *CoRR*, abs/2211.09809, 2022.

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022.

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

[20] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *CoRR*, abs/2210.02303, 2022.

[21] Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *CoRR*, abs/2204.03458, 2022.

[22] Qingqing Huang, Daniel S. Park, Tao Wang, Timo I. Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Havnø Frank, Jesse H. Engel, Quoc V. Le, William Chan, and Wei Han. Noise2music: Text-conditioned music generation with diffusion models. *CoRR*, abs/2302.03917, 2023.

[23] 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 *SIGGRAPH (Conference Paper Track)*, pages 61:1–61:10. ACM, 2022.

[24] Venkatesh Shenoy Kadandale, Juan F. Montesinos, and Gloria Haro. Vocalist: An audio-visual synchronisation model for lips and voices. In *INTERSPEECH*, pages 3128–3132. ISCA, 2022.

[25] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017.

[26] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.

[27] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *CoRR*, abs/2206.00364, 2022.

[28] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017.

[29] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

[30] Eva Lasarcyk and Jürgen Trouvain. Imitating conversational laughter with an articulatory speech synthesis. *Proceedings of the Interdisciplinary Workshop on the Phonetics of Laughter*, 04 2008.

[31] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, pages 4794–4804. IEEE, 2022.

[32] Borong Liang, Yan Pan, Zhizhi Guo, Hang Zhou, Zhibin Hong, Xiaoguang Han, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Expressive talking head generation with granular audio-visual control. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3377–3386, 2022. doi: 10.1109/CVPR52688.2022.00338. ISSN: 2575-7075.

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

[34] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021.

[35] C Niemitz. Visuelle zeichen, sprache und gehirn in der evolution des menschen—eine entgegnung auf mcfarland. *Z. Sem*, 12:323–336, 1990.

[36] Alex Pentland. *Honest signals: how they shape our world.* MIT press, 2010.

[37] Stavros Petridis, Brais Martinez, and Maja Pantic. The mahnob laughter database. *Image and Vision Computing*, 31(2):186–202, 2013.

[38] K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay 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*, pages 484–492, 2020. doi: 10.1145/3394171.3413532.

[39] Robert R Provine. *Laughter: A scientific investigation.* Penguin, 2001.

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[41] Willibald Ruch and Paul Ekman. The expressive pattern of laughter. In *Emotions, qualia, and consciousness*, pages 426–443. World Scientific, 2001.

[42] Flavio Schneider, Zhijing Jin, and Bernhard Schölkopf. Moûsai: Text-to-music generation with long-context latent diffusion. *CoRR*, abs/2301.11757, 2023.

[43] Björn Schuller, Ronald Müller, Florian Eyben, Jürgen Gast, Benedikt Hörnler, Martin Wöllmer, Gerhard Rigoll, Anja Höthker, and Hitoshi Konosu. Being bored? recognising natural interest by extensive audiovisual integration for real-life application. *Image and Vision Computing*, 27(12):1760–1774, 2009.

[44] Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized talking head synthesis. *CoRR*, abs/2301.03786, 2023.

[45] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. *CoRR*, abs/2209.14792, 2022.

[46] S. Sinha, S. Biswas, and B. Bhowmick. Identity-preserving realistic talking face generation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10, 2020.

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

[48] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[49] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*. OpenReview.net, 2021.

[50] Michal Stypulkowski, Konstantinos Vougioukas, Sen He, Maciej Zieba, Stavros Petridis, and Maja Pantic. Diffused heads: Diffusion models beat gans on talking-face generation. *CoRR*, abs/2301.03396, 2023.

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

[52] Noé Tits, Kevin El Haddad, and Thierry Dutoit. Laughter synthesis: Combining seq2seq modeling with transfer learning. In *INTERSPEECH*, pages 3401–3405. ISCA, 2020.

[53] Jérôme Urbain, Elisabetta Bevacqua, Thierry Dutoit, Alexis Moinet, Radoslaw Niewiadomski, Catherine Pelachaud, Benjamin Picart, Joëlle Tilmanne, and Johannes Wagner. The avlaughtercycle database. In *LREC*, 2010.

[54] Jérôme Urbain, Hüseyin Çakmak, and Thierry Dutoit. Automatic phonetic transcription of laughter and its application to laughter synthesis. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 153–158, 2013. doi: 10.1109/ACII.2013.32.

[55] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven facial animation with temporal gans. *BMVC*, 2018.

[56] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, pages 1–16, 2019.

[57] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. In *ICLR*. OpenReview.net, 2022.

[58] Lei Xie and Zhi-Qiang Liu. A coupled hmm approach to video-realistic speech animation. *Pattern Recognition*, 40(8):2325–2340, 2007.

[59] Eli Yamamoto, Satoshi Nakamura, and Kiyohiro Shikano. Lip movement synthesis from speech based on hidden markov models. *Speech Communication*, 26(1-2):105–115, 1998.

[60] Hani Yehia, Philip Rubin, and Eric Vatikiotis-Bateson. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1-2):23–43, 1998.

[61] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *CVPR*, pages 4176–4186. Computer Vision Foundation / IEEE, 2021.

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