Laughing Matters: Introducing Laughing-Face Generation using Diffusion Models - Extended Abstract

Antoni Bigata Casademunt Rodrigo Mira Nikita Drobyshev
Konstantinos Vougioukas Stavros Petridis Maja Pantic
Intelligent Behaviour Understanding Group (iBUG)
Imperial College London, UK
{ab4522,rs2517,k.vougioukas, stavros.petridis04,m.pantic}@ic.ac.uk, nikita.drobyshev23@gmail.com

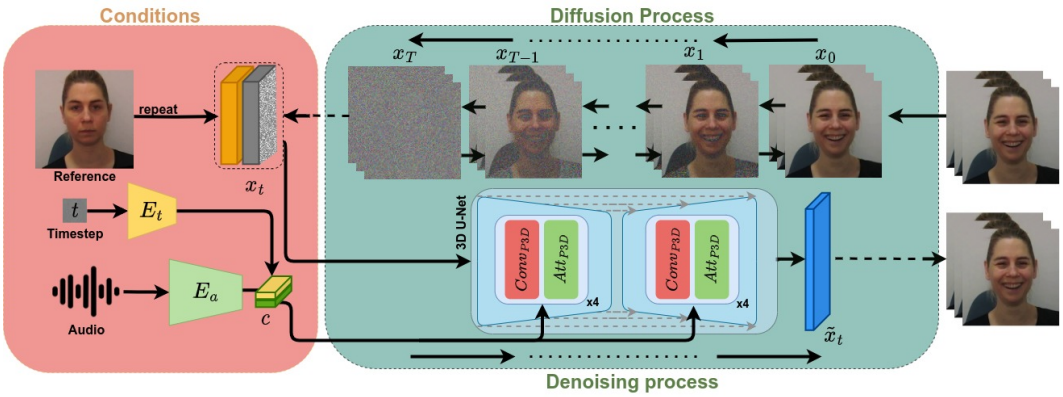

Figure 1. Overview of our proposed pipeline for laughter generation.

1. Introduction

Facial animation is vital in creating immersive and engaging experiences in various applications, such as virtual reality, movies, and human-computer interaction (HCI) [15, 19]. Current facial animation methods primarily focus on speech-driven animation, often neglecting non-verbal expressions like laughter, head nods, or blinks. These non-verbal cues convey essential contextual information and are crucial for natural dialogue [16, 18]. Laughter, in particular, is a powerful non-verbal communication medium that conveys emotions, intentions, and social relationships [4]. However, generating realistic laughter sequences is challenging due to the lack of direct correlation between laughter and lip movement, as well as the scarcity of training data.

Recent advances in speech-driven animation methods with the introduction of Generative Adversarial Networks (GANs) [5], have led to more realistic facial animations [14, 27]. Some methods even incorporate emotion control into the generation process [9, 21]. The emergence of diffusion-based generation techniques has further improved performance in this field [25]. However, current frame-based generators struggle with laughter generation due to several issues. Firstly, laughter lacks the robust audio-visual correlation seen in speech [10], making it more difficult to generate authentic audio-driven laughter sequences. Secondly, the complexity and variability of laughter, involving various muscles and facial movements, pose a substantial challenge for existing frame-based generators designed primarily for speech. Finally, the spontaneity and context-dependency of laughter make it difficult to accurately predict the timing and intensity of the speaker's facial movements.

In this paper, we propose a novel video diffusion model that leverages recent developments in video diffusion [8, 22] to generate realistic and synchronized laughing animations based on raw audio input. To address the issue of limited publicly-available audio-visual laughter corpora, we propose an ensemble of existing datasets for training and evaluation. We employ metrics from existing video generation works and design a novel metric specifically tailored for laughter generation to assess the quality of our results.

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. Methodology

We summarize our approach in Figure 1. We use diffusion models [6, 23, 24] which are generative models that synthesize data by iteratively removing Gaussian noise. We follow the approach of Karras *et al*. [11], using Heun's method for efficient noise prediction. The diffusion process uses a noise schedule with standard deviation σ_t and a Gaussian distribution ($\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$). We train a model D_θ to predict the original sample from a noisy version.

We build upon the factorized space-time U-Net architecture of Ho *et al*. [8]. Our input is a video ($x_t \in \mathbb{R}^{B\times C\times F\times H\times W}$), conditioned on a reference frame (x') and audio (a). We use an audio (E_a) and timestep encoder (E_t). Pseudo-3D convolution and attention layers [22] improve efficiency. Audio and timestep information are fed into each U-Net ResBlock.

We address limited laughter datasets with the following techniques:

Augmentation regularization. We use geometric transformations to prevent overfitting [11, 12].

Classifier-free guidance (CFG). This improves sample-condition alignment [7].

Longer sequence inference. We train on shorter video segments (16 frames) as the duration of laughter is usually short but generate arbitrary lengths during inference.

3. Experiments

Datasets and Evaluation Metrics We use four laughter-containing datasets: MAHNOB [17], AVLaughterCycle [26], AVIC [20], and SAL [3]. We focus on laughter-specific segments and create an 80/10/10 (train/validation/test) split with no speaker overlap. Data was preprocessed with face alignment, normalization, a 16kHz audio sampling rate, and a 25fps video frame rate.

We employ several metrics to assess our model's output:

Reconstruction Quality. Frechet Inception Distance (FID) and Structural Similarity Index (SSIM) measure the similarity between generated and real images.

Visual Quality, Coherence and Diversity: Frechet Video Distance (FVD) offers a more holistic evaluation of video outputs.

Laughter Authenticity: We train a Laughter Classifier (LC) on MAHNOB data (fine-tuning a MViTv2 backbone [13]) to distinguish between laughter and speech. This LC is applied to generated videos, with high laughter classification rates indicating realistic laughter synthesis.

Audio Encoder	FVD ↓	FID ↓	SSIM ↑	LC ↑ (%)
SDA [27]	169.48	55.07	0.318	68.21
WavLM [1]	136.76	46.01	0.312	54.21
Mel-spectrograms	124.81	47.74	0.320	83.52
BEATs [2]	**111.95**	**45.69**	**0.371**	**96.52**

Table 1. Ablation study on the audio encoder.

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 2. Ablation study on the training improvements.

Results As this is the first work on audio-driven laughter generation, we compare against speech-driven animation methods retrained on laughter data: Diffused Heads [25], SDA [27], and EAMM [9]. We also include pre-trained models like MakeItTalk [29] and PC-AVS [28]. Table 3 highlights the limitations of pre-trained models and the improvements gained by retraining on laughter-specific data.

Our method consistently outperforms others in visual quality and laughter authenticity. We attribute this to **3D Architecture** that enables modelling longer audio context, crucial for laughter's weaker audiovisual correlation compared to speech. To the **laughter-specific audio encoder**, BEATs [2], which outperforms speech-focused encoders (See Table 1). And finally, to the **training improvements** we used to compensate for limited laughter datasets (see Table 2). We further confirm our model's superiority through a Mean Opinion Score (MOS) test. Participants rated videos on a scale of 1 (clearly artificial) to 5 (highly realistic). Our model significantly improves upon competitors, even though ground-truth videos received relatively low scores, underscoring the difficulty of judging laughter realism.

4. 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 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

Model	FVD ↓	FID ↓	SSIM ↑	LC ↑ (%)	MOS ↑
Pre-trained					
Diffused Heads [25]	149.51	49.36	0.236	80.70	-
SDA [27]	594.32	111.89	0.053	13.85	-
EAMM [9]	391.62	71.71	0.094	16.67	-
PC-AVS [28]	1164.49	175.99	0.004	53.91	-
MakeItTalk [29]	196.89	49.08	0.262	72.50	1.94±1.12
Re-trained					
Diffused Heads [25]	152.30	67.46	0.232	94.09	2.45±1.22
SDA [27]	696.33	124.52	0.040	85.13	-
EAMM [9]	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 3. Comparative performance of the proposed methods against pre-trained and re-trained models. The best result is in **bold**.

be promising to extend our model to cover other non-verbal cues and create a comprehensive facial animation model that can animate all verbal and non-verbal cues in natural speech.

References

[1] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, M. Zeng, X. Yu, and F. 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. 2

[2] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, and F. Wei. Beats: Audio pre-training with acoustic tokenizers. *CoRR*, 12 2022. 2

[3] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amir, and D. 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. 2

[4] P. Glenn. *Laughter in interaction*, volume 18. Cambridge University Press, 2003. 1

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[6] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[7] J. Ho and T. Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022. 2

[8] J. Ho, T. Salimans, A. A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet. Video diffusion models. *CoRR*, abs/2204.03458, 2022. 1, 2

[9] X. Ji, H. Zhou, K. Wang, Q. Wu, W. Wu, F. Xu, and X. 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. 1, 2, 3

[10] V. S. Kadandale, J. F. Montesinos, and G. Haro. Vocalist: An audio-visual synchronisation model for lips and voices. In *INTERSPEECH*, pages 3128–3132. ISCA, 2022. 1

[11] T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models. *CoRR*, abs/2206.00364, 2022. 2

[12] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020. 2

[13] Y. Li, C. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, pages 4794–4804. IEEE, 2022. 2

[14] B. Liang, Y. Pan, Z. Guo, H. Zhou, Z. Hong, X. Han, J. Han, J. Liu, E. Ding, and J. 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. ISSN: 2575-7075. 1

[15] C. Niemitz. Visuelle zeichen, sprache und gehirn in der evolution des menschen—eine entgegnung auf mcfarland. *Z. Sem*, 12:323–336, 1990. 1

[16] A. Pentland. *Honest signals: how they shape our world*. MIT press, 2010. 1

[17] S. Petridis, B. Martinez, and M. Pantic. The mahnob laughter database. *Image and Vision Computing*, 31(2):186–202, 2013. 2

[18] R. R. Provine. *Laughter: A scientific investigation*. Penguin, 2001. 1

[19] W. Ruch and P. Ekman. The expressive pattern of laughter. In *Emotions, qualia, and consciousness*, pages 426–443. World Scientific, 2001. 1

[20] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu. Being bored? recognising natural interest by extensive audiovisual integration for real-life application. *Image and Vision Computing*, 27(12):1760–1774, 2009. 2

[21] S. Shen, W. Zhao, Z. Meng, W. Li, Z. Zhu, J. Zhou, and J. Lu. Difftalk: Crafting diffusion models for generalized talking head synthesis. *CoRR*, abs/2301.03786, 2023. 1

[22] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, and Y. Taigman. Make-a-video: Text-to-video generation without text-video data. *CoRR*, abs/2209.14792, 2022. 1, 2

[23] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2

[24] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 2

[25] M. Stypulkowski, K. Vougioukas, S. He, M. Zieba, S. Petridis, and M. Pantic. Diffused heads: Diffusion models beat gans on talking-face generation. *CoRR*, abs/2301.03396, 2023. 1, 2, 3

[26] J. Urbain, E. Bevacqua, T. Dutoit, A. Moinet, R. Niewiadomski, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner. The avlaughtercycle database. In *LREC*, 2010. 2

[27] K. Vougioukas, S. Petridis, and M. Pantic. End-to-end speech-driven facial animation with temporal gans. *BMVC*,

2018. 1, 2, 3

[28] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *CVPR*, pages 4176–4186. Computer Vision Foundation / IEEE, 2021. 2, 3

[29] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6):1–15, 2020. 2, 3