Supplementary Material for: "Laughing Matters: Introducing Audio-Driven Laughing-Face Generation with 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

1 Datasets

Training split

The models are trained and evaluated on MAHNOB, AVLaughterCycle, AVIC and SAL. The speakers used for training, validation and testing are shown in Table 1.

Dataset	Training	Validation	Testing
AVLaughterCycle	11,13,18,5,6,7	16	14
Mahnob	1,2,4,6,7,8,9,11,13,15,17,19,20,22,23,24,25	5,16	3,14,21
AVIC	4,5,6,8,9,10,12,13,26,27,30,31,32,33,34,35	16,36	15,28,29
SAL	Alex, Donn, Gary, Liam, Mlind, Nicol, Ruth, Shar, Ed, Ian, Rod	Mart, GHill	Nol, Alis

Table 1: Speaker IDs for training, validation and test sets.

2 Laughter Classifier

We introduce the Laughter Classifier as part of our evaluation methodology. This not only highlights the limitations of pre-trained speech-driven animation methods, but also demonstrates the capabilities of our model in generating realistic laughter sequences. The model processes video inputs and produces a single logit output to classify whether the individual in the video is speaking or laughing.

2.1 Architecture

The architecture of the system is presented in Fig. 1. We employ a Multiscale Vision Transformers (MViT) [2] backbone with two linear layers and a dropout layer with a dropout probability set at 0.2. The MViT model, pre-trained on the Kinetics 400 dataset, reaches a top-5 accuracy of 94.665 %.

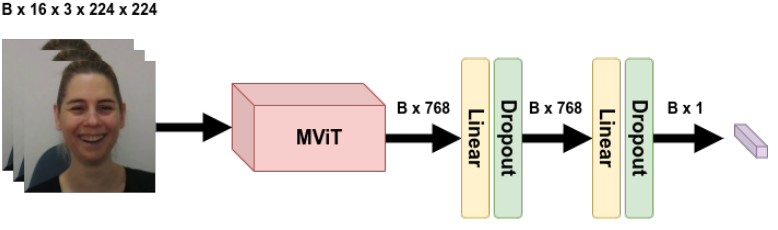

Figure 1: The architecture used for the Laughter Classifier. The batch size is denoted as B.

2.2 Training

Our model was trained using the MAHNOB training set for laughter combined with additional speech videos for each speaker, given that MAHNOB also provides speech data. During the training process, either laughter or a speech video is fed into the model with equal probability. We train with the AdamW optimizer, configured with a learning rate of 1×10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Binary Cross Entropy is employed as the loss function.

3 Study on CFG scale

We investigated the effects of the CFG scale on the output of our diffusion model. The CFG scale is a parameter that controls how closely the generated image adheres to the user's condition. We find that a higher CFG scale value resulted in images that were more faithful to the condition, but also had more artefacts. A lower CFG scale value resulted in images with less noise, but they were also less faithful to the condition. We show in Fig. 2 a comparison between FID and FVD for scales between 0 and 12 and found CFG=1 to be the optimal tradeoff in our case.

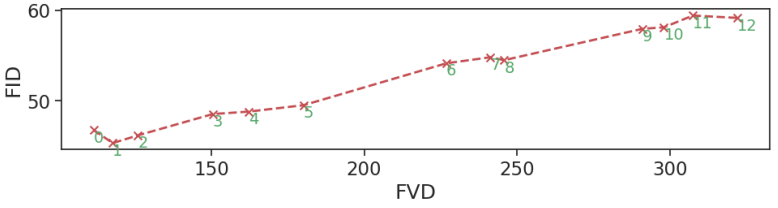

Figure 2: CFG analysis by comparing FID and FVD for different scales

4 User study

Since quality metrics do not always align with human perception, we conduct a Mean Opinion Score evaluation. In this study, participants rate videos from different models on a scale of 1 to 5, with 1 indicating that the video appears clearly artificial, and 5 signifying that it is highly realistic and indistinguishable from genuine laughter. Each participant views a minimum of 12 videos, with the option to extend their participation up to 60 videos. We have gathered 72 responses in total, averaging 23 videos per individual participant. The distribution of responses for each model is depicted in Fig. 3 and Fig. 4.

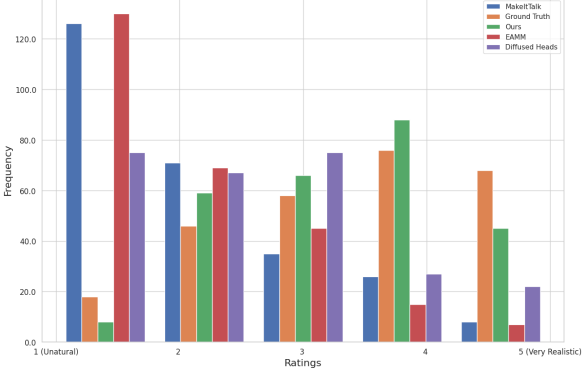

Figure 3: Histogram of the user study rating for our model compared to MakeItTalk [5], EAMM [1], Diffused Heads [4] and Ground Truth.

5 Limitations

Throughout the evaluation process, we identified potential issues related to long-term generation, attributed to the autoregressive nature of the process and the limited data availability. Generating sequences longer than 2 seconds (or 50 frames) resulted in degraded quality. An example of such failure case for longer generation sequences can be observed in Fig. 5. In our specific use case, this is not a major concern, as laughter is typically brief. For instance, the average laughter duration in the MAHNOB dataset is 1.56 seconds, suggesting that our model can handle the majority of laughter episodes. However, this highlights a valuable direction for future research. One potential solution could involve conditioning the model via

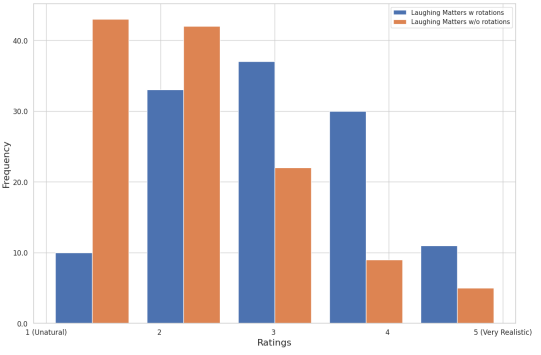

Figure 4: Histogram of the user study rating for our model with and without head rotations.

an additional, unchanging *identity frame*. This strategy could provide the model with enough information to maintain consistent quality throughout the autoregressive generation process.

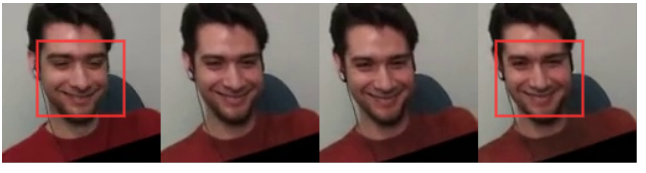

Figure 5: Failure case of our network.

6 Videos

For a more comprehensive understanding of our findings, we encourage readers to examine the side-by-side comparison video provided as part of the supplementary material. Within the same video, we also demonstrate a comparison between voiced and unvoiced laughter, highlighting our model's ability to generate both types. Voiced laughter is characterized by its harmonically rich, vowel-like sound, accompanied by measurable periodicity in vocal fold vibration. On the other hand, unvoiced laughter is produced through a noisy exhalation from either the nose or mouth, without involvement of the vocal folds. This distinction is crucial as research has shown that these two forms of laughter serve different functions in social interactions [3].

References

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.